



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Surveillance Metrics Sensitivity Study

R. Bierbaum, M. Hamada, A. Robertson

November 4, 2011

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

SANDIA REPORT

SAND2011-6763

Unlimited Release

Printed September 2011

Surveillance Metrics Sensitivity Study

Rene L. Bierbaum

Michael S. Hamada

Alix A. Robertson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.



Sandia National Laboratories



Lawrence Livermore National Laboratory



Los Alamos National Laboratory

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.osti.gov/bridge>

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>



Surveillance Metrics Sensitivity Study

Rene L. Bierbaum
Sandia National Laboratories
Livermore, CA 94550

Michael S. Hamada
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Alix A. Robertson
Lawrence Livermore National Laboratory
Livermore, CA 94550

Abstract

In September of 2009, a Tri-Lab team was formed to develop a set of metrics relating to the NNSA nuclear weapon surveillance program. The purpose of the metrics was to develop a more quantitative and/or qualitative metric(s) describing the results of realized or non-realized surveillance activities on our confidence in reporting reliability and assessing the stockpile. As a part of this effort, a statistical sub-team investigated various techniques and developed a complementary set of statistical metrics that could serve as a foundation for characterizing aspects of meeting the surveillance program objectives. The metrics are a combination of tolerance limit calculations and power calculations, intending to answer level-of-confidence type questions with respect to the ability to detect certain undesirable behaviors (catastrophic defects, margin insufficiency defects, and deviations from a model). Note that the metrics are not intended to gauge product performance but instead the adequacy of surveillance.

This report gives a short description of four metrics types that were explored and the results of a sensitivity study conducted to investigate their behavior for various inputs. The results of the sensitivity study can be used to set the risk parameters that specify the level of stockpile problem that the surveillance program should be addressing.

ACKNOWLEDGMENTS

The authors would like to thank the following for their insights, lively discussion, and feedback on this work: Surveillance Metrics Team Lead Bill McLean (Lawrence Livermore National Laboratory), Aparna Huzurbazar and Earl Whitney (Los Alamos National Laboratory), Rena Zurn (Sandia National Laboratories) and Kathleen Diegert and Ron Streit (NNSA points-of contact).

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

This work was performed under the auspices of the Los Alamos National Laboratory, an affirmative action/equal opportunity employer, operated by the Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396.

CONTENTS

Acknowledgments.....	4
1. Introduction.....	7
2. Background and Scope of the Study	7
3. Presentation of Results	10
3.1. Metric 1a/1b Discussion	10
3.2. Metric 2a/2b/2c Background.....	14
3.3. Metric 2a Discussion	14
3.4. Metric 2a Discussion (Normal Population Distribution)	16
3.5. Metric 2a Discussion (Lognormal Population Distribution)	20
3.6. Metric 2b/2c Discussion	23
3.7. Metric 3a/3b Discussion	32
3.8. Metric 4 Discussion (Linear Trend)	35
3.9. Metric 4 Discussion (Non-Linear Trend)	41
4. Conclusions and Recommendations.....	43
Appendix A: R Code.....	44
Metric 1a.....	44
Metric 1b.....	47
Metric 2a.....	49
Metric 2b/2c.....	56
Metric 3a/3b.....	61
Metric 4	63
Appendix B: Metric 1a and 1b Sensitivity Study Graphs	71
Appendix C: Metric 2a Sensitivity Study Graphs, Normal Distribution	72
Appendix D: Metric 2a Sensitivity Study Graphs, Log-Normal Distribution.....	80
Appendix E: Metric 2b Sensitivity Study Graphs	88
Appendix F: Metric 2c Sensitivity Study Graphs.....	93
Appendix G: Metric 3a/3b Sensitivity Study Graphs	96
Appendix H: Metric 4 Sensitivity Study Graphs, Linear Trend.....	97
Appendix I: Metric 4 Sensitivity Study Graphs, Non-Linear Trend	107

TABLES

Table 1: Sensitivity Study Work Completed.....	9
Table 2: Metric 1a/b Number of Samples (Population of 500)	12
Table 3: Metric 1b by Annual Sample Rate and Defect Size	13
Table 4: Metric 2a Median Number of Samples (Target: 0.9 within spec).....	18
Table 5: Metric 2a by True (Target: 0.95 within spec)	20
Table 6: Metric 2b for End=0.95 (Target: 0.9 within spec)	27
Table 7: Metric 2b for End=0.99 (Target: 0.9 within spec)	28
Table 8: Matrix H and Tolerance Factor k	29
Table 9: Metric 3a by Size of Shift in Mean (in Standard Deviations, SD).....	34
Table 10: Metric 3b by Size of Increase in Standard Deviation (in Standard Deviations, SD)	35
Table 11: Metric 4, Sampling over 30 Years (Start=0.9999, Start+15 Years=0.95).....	40
Table 12: Metric 4, Sampling over 10 Years (Start=0.9999, Start+15 Years=0.95).....	40

FIGURES

Figure 1: Representative Graphs for Metric 1a and 1b	12
Figure 2: Metric 2a Summary Graphs (Normal Distribution)	17
Figure 3: Metric 2a when True=Target.....	19
Figure 4: Non-Normal Distributions for Metric 2a: Normal vs. LN Same Variance	21
Figure 5: Non-Normal Distributions for Metric 2a: Normal vs. LN Larger Variance	22
Figure 6: Comparison of Metric 2a for Normal and Lognormal Distributions	23
Figure 7: Comparison of Metric 2b For Different Sample Period, Different Start but Same End(per).....	26
Figure 8: Comparison of Metric 2a and 2b	27
Figure 9: Representative Graphs for Metric 2c	28
Figure 10: Tolerance Limits for Simulated Data Sets at 3/yr Over 10-Years	30
Figure 11: Tolerance Limits for Four Simulated Data Sets at 1/yr Over 30-Years.....	31
Figure 12: Metric 3a/3b Summary Graphs.....	34
Figure 13: Representative Graph for Metric 4, Linear Trend.....	39
Figure 14: Example of Linear and Non-Linear Trends.....	41
Figure 15: Representative Graph for Metric 4, Comparison of Linear and Non-Linear Trends...	42

1. INTRODUCTION

In September of 2009, a Tri-Lab team was formed to develop a set of metrics relating to the NNSA nuclear weapon surveillance program. The purpose of the metrics was to develop a more quantitative and/or qualitative metric(s) describing the results of realized or non-realized surveillance activities on our confidence in reporting reliability and assessing the stockpile. As a part of this effort, a statistical sub-team investigated various techniques and developed a complementary set of statistical metrics that could serve as a foundation for characterizing aspects of meeting the surveillance program objectives. The metrics are a combination of tolerance limit calculations and power calculations, intending to answer level-of-confidence type questions with respect to the ability to detect certain undesirable behaviors (catastrophic defects, margin insufficiency defects, and deviations from a model). Note that the metrics are not intended to gauge product performance but instead the adequacy of surveillance.

This report gives a short description of four metrics types that were explored and the results of a sensitivity study conducted to investigate their behavior for various inputs. The results of the sensitivity study can be used to set the risk parameters that specify the level of stockpile problem that the surveillance program should be addressing.

2. BACKGROUND AND SCOPE OF THE STUDY

The objective of the surveillance metrics is to provide a statistical measure of certain aspects of the surveillance program relating to data collection. Hence the inputs of these metrics all relate to some aspect of the surveillance program, with the output being a measure of that program's ability to identify selected deleterious behaviors such as defects or trends. However each of the surveillance metrics requires one or more decisions relating to the amount of risk one is willing to take; for example, the size of the defect in the stockpile. Further, the tolerance limit metrics depend upon the initial margin of the product. And finally, there are assumptions associated with some of the metrics (e.g., linearity, normality) that may not be met.

As can be seen from this, the scope of the sensitivity study is potentially very broad. For this exercise, there were four general areas of focus that were identified prior to the study for exploration:

- Parameter values
- Trend scenarios
- Sampling profiles
- Other data considerations

Further, three classes of parameters were identified for exploration:

- characteristics of the surveillance program
- characteristics of the product and measurement system
- risk characteristics

In general, linear trends were examined but some other trend scenarios based upon past experiences with real data were explored for one of the metrics. This will be discussed later in the paper.

Some of the metrics also depend upon how the samples are allocated over time, in addition to how many there are. The intent was thus to examine cases where samples were allocated at regular intervals as well as in “clumps”. Again, this was explored only to a limited extent to date but with some interesting results.

The other data considerations listed below were also considered to be important but were not fully explored. In particular, non-homogeneity is known to exist in the stockpile and poses unique surveillance challenges but its implications for the variables data metrics are not yet fully explored.

- A. Sub-populations
- B. Non-normal data (one case explored)
- C. Other data anecdotes?
- D. Measurement error issues: presumed to be convolved with the data

Table 1 summarizes the sensitivity study plan that the authors completed for this report. The listed metrics are defined in the body of the report. R code¹ was written to do the calculations. The code is included in Appendix A.

¹ R is a free software environment for statistical computing and graphics. It can be downloaded at <http://www.r-project.org>.

Table 1: Sensitivity Study Work Completed

		Metric 1a	Metric 1b	Metric 2a	Metric 4	Metric 2b/c	Metric 3a/b
Characteristics of the Surveillance Program	Total Samples	1-50 50-250		1-50			
	Sample History		1-100 1-20		Evenly distributed vs. split 50:50, 80:20 beginning and end; Over 10, 30 years N=1-100	Evenly distributed, Over 10, 30 years N=1-100	1-50
Characteristics of the Product and Measurement System	Population size	100, 500, infinite	100, 500, infinite				
	Non-normality			Normal LN (same SD, 5x SD as normal)			
	Initial (or static) probability within spec.			.8, .9, .95, .99, .995, .999, .99999	.8, .9, .95, .99, .999, .9999, .999999	.8, .9, .95, .99, .999, .9999, .999999	
	SNR (slope/s _{slope})				Determined by initial and final margin and work back to give .01, .05, .1	Ending proportion .0001, .001, .01, .05, .1, .25	
Risk Characteristics	Defect Size	.01, .05, .1, .25	.01, .05, .1, .25	.01, .05, .1, .25	.0001, .001, .01, .05, .1, .25	.1	
	Lookback Time		4 yrs*		From time zero		4 yrs*
	Predictive Time				15 years from today	(a) today (b) 15 years from today	
	False Positive Rate				5%		5%
	Size of Effect						(a) 1, 2, 3 sigma (b) 1.5, 2, 3 sigma

*This does not affect the sensitivity calculation, but, of course, the number of samples and hence power will depend on the number of years to detect a change.

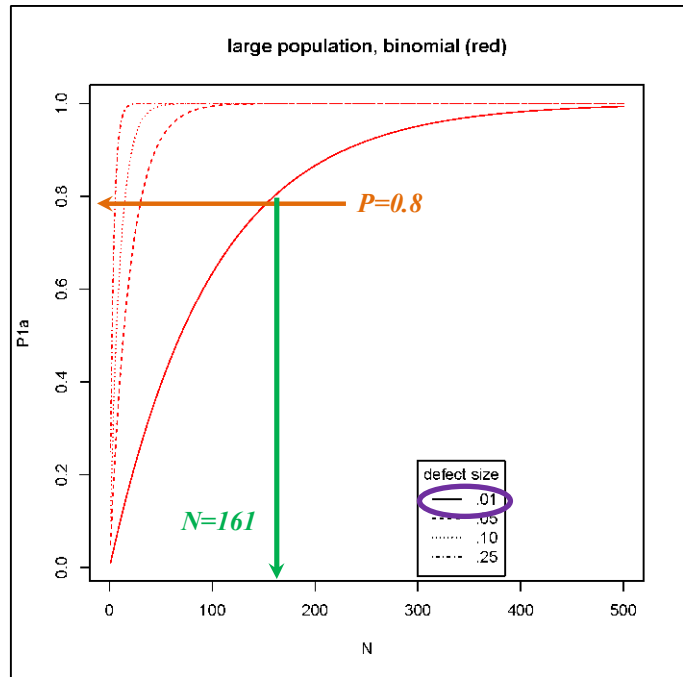
3. PRESENTATION OF RESULTS

For each of the metrics, the following will be presented:

- A brief description of the metric and its role
- One or more representative graphs of results, illustrating the general metric behavior
- One or more representative tables of results which help to highlight the relationship between N (the number of samples) and the metric value under various conditions
- A discussion of general observations, insights, and dependencies

An example for Metric 1a is shown below to illustrate how to interpret the graphs.

	P_{1a} level			
Defect rate	0.6	0.7	0.8	0.9
0.01	92	120	161	230
0.05	18	24	32	45
0.1	9	12	16	22
0.25	4	5	6	9



The x-axis of the plot is N, the number of samples (the input characterizing the surveillance program in this case). There are separate curves drawn to show the dependence on risk level (in this case, defect size). Other assumptions are noted in the heading of the graphs (e.g., the “large population” case). The y-axis is the metric value calculated for each surveillance program input, risk level, and other assumptions.

3.1. Metric 1a/1b Discussion

The purpose of Metric 1a/1b is to assess the power to detect defective units. The power to detect defective units is addressed by two sub-metrics:

P_{1a} : The probability of identifying a static defect in stockpile.

P_{1b} : The probability of identifying a latent defect in stockpile.

For large populations, one can use the binomial distribution to calculate Metric 1a/1b. If N_{total} is the total number of samples and p is the proportion in the population with a problem,

$$P_{1a} = 1 - (1 - p)^{N_{total}} \quad (1)$$

If N_{xyeas} is the total number of samples in the last x years and p is the proportion in the population with a problem,

$$P_{1b} = 1 - (1 - p)^{N_{xyeas}} \quad (2)$$

For small populations, we use the hypergeometric distribution. If $N_{population}$ is the size of the population and N_{total} is the total number of samples and p is the proportion in the population with a problem,

$$P_{1a} = 1 - \frac{\binom{N_{population} p}{0} \binom{N_{population} - N_{population} p}{N_{total}}}{\binom{N_{population}}{N_{total}}} \quad (3)$$

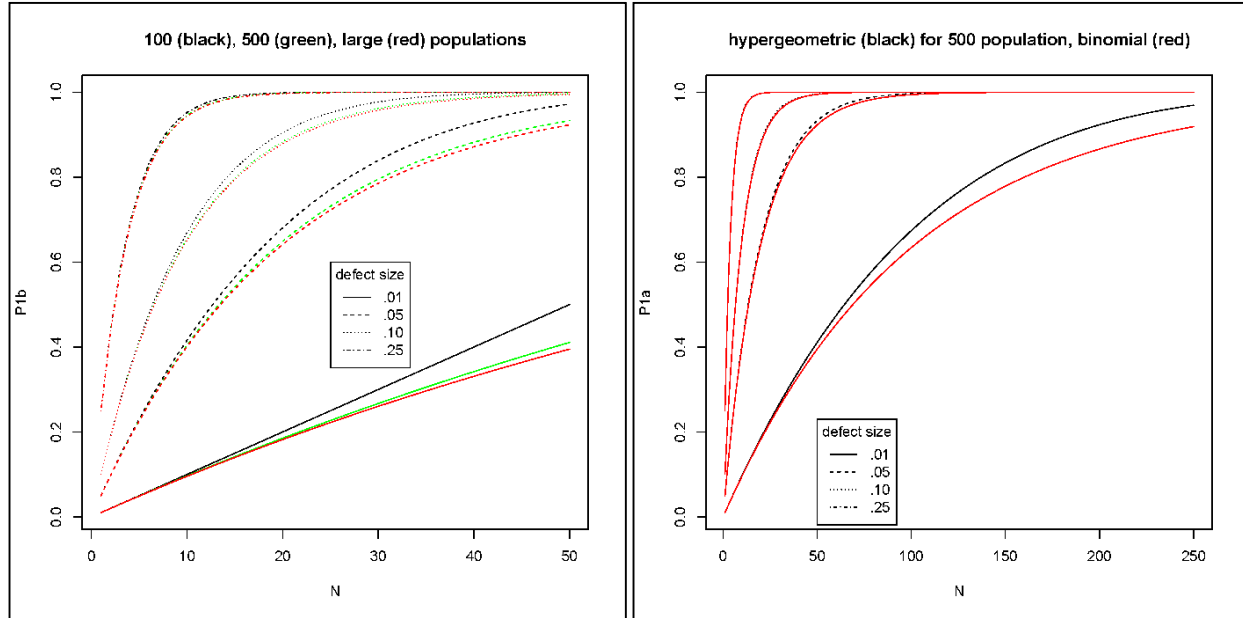
If N_{xyeas} is the total number of samples in the last x years and p is the proportion in the population with a problem,

$$P_{1b} = 1 - \frac{\binom{N_{population} p}{0} \binom{N_{population} - N_{population} p}{N_{xyeas}}}{\binom{N_{population}}{N_{xyeas}}} \quad (4)$$

Because of their close relationship, the results for Metrics 1a and 1b will be discussed together. In each case we use N to represent the total number of samples that are counted to detect the static or latent defect.

The graphs below show the general behavior of the metric.

Figure 1: Representative Graphs for Metric 1a and 1b
N=1-50 N=1-250



The left graph shows metric calculations for sample sizes ranging from one to 50, and defect sizes of one, five, ten, and 25 percent. Two sets of curves show exact hypergeometric calculations for sampling from populations of 100 (black) and 500 (green). A third set of curves show approximate, binomial calculations for sampling from an arbitrarily large population (red).

The right graph compares hypergeometric (black) and binomial (red) calculations for sampling from a total population of 500, with sample sizes of one to 250. The table shows the results for sampling from a total population of 500 in terms of the number of samples needed to achieve metric levels of 0.6, 0.7, 0.8, and 0.9.

See Appendix B for the entire set of graphs generated for Metrics 1a and 1b.

Table 2: Metric 1a/b Number of Samples (Population of 500)

	P_{1a} or P_{1b} level			
Defect size	0.6	0.7	0.8	0.9
0.01	84	107	138	184
0.05	18	23	31	43
0.1	9	12	16	22
0.25	4	5	6	8

General observations include the following:

- There is a strong dependence upon the risk level (defect size) selected. More samples are required to detect smaller defect sizes. This conforms to one's intuition.

- There is a strong dependence upon the number of samples, N . This dependence is approximately linear for small sample sizes. However there is a knee in the curve above which there is relatively little improvement in the metric once N reaches a certain value. The location of the knee in the curve depends upon the defect size and also depends on the population size.
- For population size of 500 and defect sizes of five percent or greater, the binomial distribution offers an acceptable approximation, although certainly the hypergeometric distribution calculations are easy with Excel or other tools.
- The definition of P_{1b} includes a time window to gauge data currency. The calculation is independent of this choice, but it will obviously affect the value of N which is a critical input. The table below shows the value of P_{1b} for various annual sample rates and a 4-year window. It assumes an arbitrarily large population. For small annual sample rates, this metric will have a low value. For example, even for a 10% defect size the metric value is only 0.34 for a sample rate of one per year. Even so, this underscores an important fact about the surveillance program. If the population is not evolving homogeneously (i.e., only a subset is defective at a given point in time), one needs several samples per year to have a chance of detecting emerging defectiveness in a timely fashion.
- The choice of time window for P_{1b} poses a tradeoff. Experience shows that latent defects can show up in a relatively short period of time, so the time window should not be too long or the metric will be misleading in its assessment of data currency. On the other hand, the surveillance program experiences ebbs and flows in testing and we do not want the metric artificially “bouncing” due simply to logistical issues. Somewhere in the 3-5 year range should provide a reasonable compromise for the time window, although certainly other choices can be made for selected datasets based upon other considerations.

Table 3: Metric 1b by Annual Sample Rate and Defect Size

Annual Sample Rate	4-Year Total	P_{1b} (.01 defect)	P_{1b} (.05 defect)	P_{1b} (.1 defect)	P_{1b} (.25 defect)
1	4	0.04	0.19	0.34	0.68
3	12	0.11	0.46	0.72	0.97
5	20	0.18	0.64	0.88	1.00
7	28	0.25	0.76	0.95	1.00
9	36	0.30	0.84	0.98	1.00
11	44	0.36	0.90	0.99	1.00

- One important assumption for P_{1a} and P_{1b} that was not explored in the sensitivity study is that it is presumed that any defect present in a sample will be detected by the testing that is performed on that sample. That is less a mathematical question than an engineering question, yet it greatly affects the outcome. In most cases, the most expedient means of dealing with this is by sub-setting the data appropriately (e.g., including in N only cold temperature tests if there is a failure mechanism known to only be manifested at cold temperature). However if one is searching for unknown mechanisms, it is impossible to anticipate what sub-setting might be appropriate. It should thus always be borne in mind that P_{1a} and P_{1b} may be misleadingly high for particular mechanisms that are not detectable in all tests. Continuous engineering review of test efficacy and diversity is as

important as continuous review of test quantities in understanding the state of the stockpile.

3.2. Metric 2a/2b/2c Background

The purpose of Metric 2a/2b/2c is to reflect the availability of sufficient test data to determine whether a threshold is met, not met, or there is insufficient data to determine whether the threshold is met or not met. The metric is calculated by using a statistical tolerance limit approach to find the confidence that a specified proportion of the population meets the performance threshold. In the case of a lower spec limit, the metric is the confidence associated with the lowest one-sided tolerance limit that still exceeds the spec limit. Similarly, in the case of an upper spec limit, the metric is the confidence associated with the highest one-sided tolerance limit that still does not exceed the limit.

Confidence in margins is addressed by three sub-metrics:

P_{2a} : The confidence that a proportion of a specified performance variable does (or does not) exceed a threshold value today, when margins are assumed not to be changing.

P_{2b} : The confidence that a proportion of a specified performance variable does (or does not) exceed a threshold value today, when data are trended to allow for the possibility that margins may be changing.

P_{2c} : The confidence that a proportion of a specified performance variable does (or does not) exceed a threshold value at a specified time in the future, when data are trended to allow for the possibility that margins may be changing.

Each of these submetrics will be addressed in turn.

3.3. Metric 2a Discussion

Metric 2a is the confidence that a proportion of the population does (or does not) exceed a threshold value today, if margins are assumed not changing. We calculate Metric 2a by finding the confidence associated with the lowest static one-sided tolerance limit that still exceeds the threshold, or the highest static one-sided tolerance limit that still does not exceed the threshold.

The static, one-sided lower tolerance limit (Y_L) and upper tolerance limit (Y_U) are calculated from a series of N data points as:

$$Y_L = \bar{Y} - ks \quad (5)$$

$$Y_U = \bar{Y} + ks \quad (6)$$

in which k is the tolerance factor, \bar{Y} is the sample mean, and s is the sample standard deviation.

The exact analytical expression for the tolerance factor is given by:

$$k = \frac{T_{\nu, \delta, \gamma}^{-1}}{\sqrt{N}} \quad (7)$$

in which $T_{\nu, \delta, \gamma}^{-1}$ is the $100\gamma\%$ percentile of the non-central t-distribution with $\nu = N - 1$ degrees of freedom and non-centrality parameter δ . The non-centrality parameter is given by:

$$\delta = \sqrt{N} z_P \quad (8)$$

in which z_P is the $100P\%$ percentile of the standard normal distribution, and P is the target conforming proportion of the population that meets the threshold.

In the case of a lower performance limit, we wish to set $Y_L = L$, the performance limit. Solving for confidence:

$$P_{2a} = T_{\nu, \delta} \left(\frac{\bar{Y} - L}{\frac{S}{\sqrt{N}}} \right) \quad (9)$$

in which $T_{\nu, \delta}$ is the cumulative distribution function of the non-central t-distribution with $\nu = N - 1$ degrees of freedom and non-centrality parameter δ . \bar{Y} and S are the sample mean and standard deviation of the data.

Similarly, for an upper performance limit, the metric is calculated from:

$$P_{2a} = T_{\nu, \delta} \left(\frac{L - \bar{Y}}{\frac{S}{\sqrt{N}}} \right) \quad (10)$$

In the sensitivity study, three key parameters were varied:

1. **Target conforming proportion, “target” (P):** First of all, the target conforming proportion had to be specified. This is referred to as the “target” in the graphs and signifies the particular proportion of the distribution, P , lying inside of the specification limit that one wishes to assure with some confidence. The target values selected for the sensitivity study were 0.75, 0.9, 0.95, and 0.99.
2. **True conforming proportion, “true”:** Metric 2a also depends upon the true proportion of the population lying inside of the specification limit. This is referred to as “true” in the graphs and signifies the true conforming proportion. The true values selected for the sensitivity study were 0.8, 0.9, 0.95, 0.99, 0.995, 0.999, and 0.99999. For the purpose of simulation, the true conforming proportion was set by drawing random samples from a standard normal distribution and setting the performance limit equal to the appropriate percentile of the standard normal.

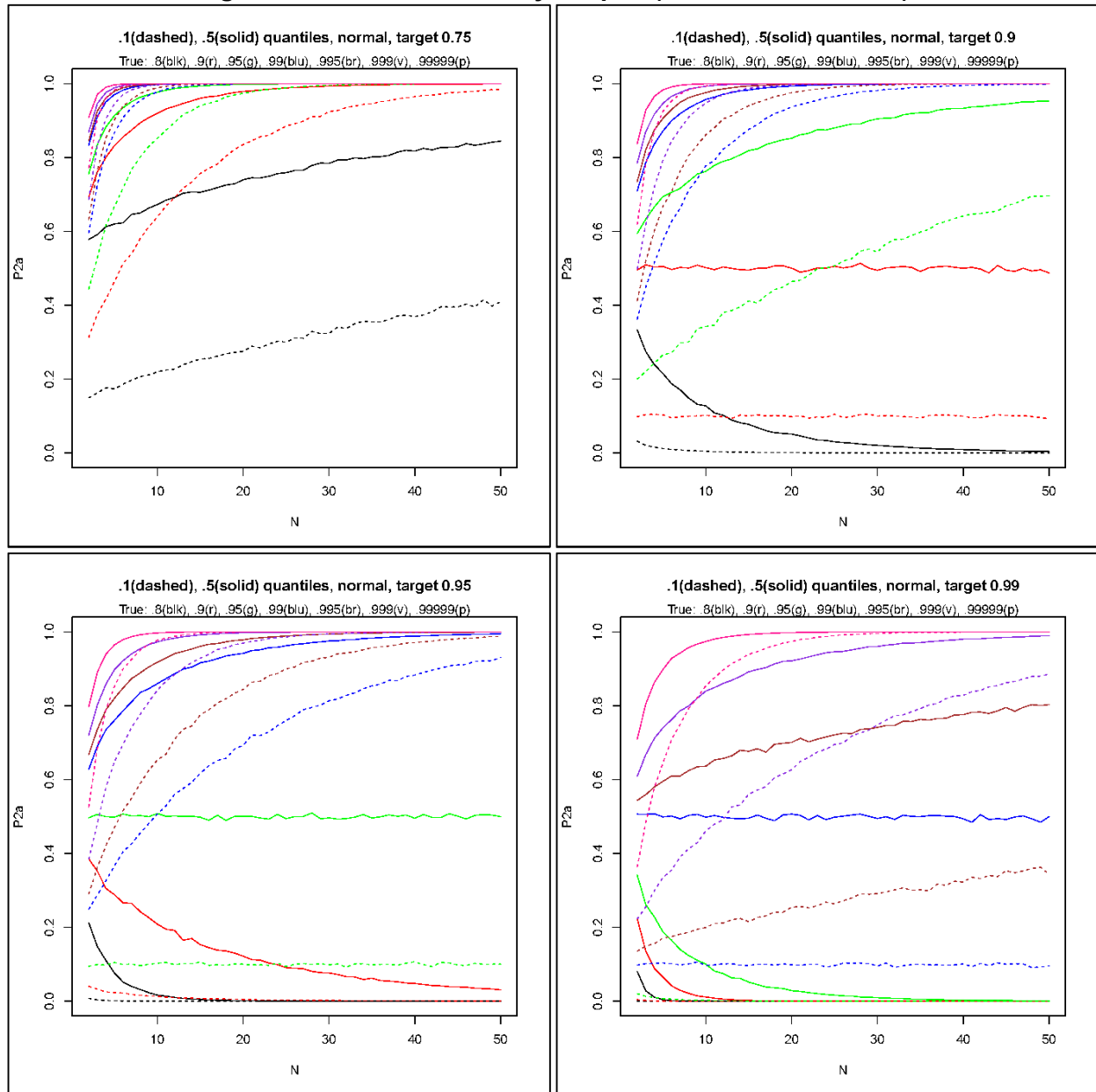
3. **True distribution:** Metric 2a calculates confidence in margin based on the assumptions that unit-to-unit variation is normal in distribution. A study was done for Metric 2a to see what impact non-normality might have. Two non-normal distributions were chosen to compare to the results for a normal distribution: a lognormal distribution with the same variance as the normal distribution and a lognormal distribution with a variance approximately five times that of the normal distribution.

Note that Metric 2a depends not just on the values of “true” and “target”, but also on sampling variability. This random variation is inevitable because the true conforming proportion in the population is not known, but is estimated from sample data. In order to understand the uncertainty in the value of P_{2a} introduced by this sampling variability, 10,000 simulations were run for each set of parameters. Quantiles of P_{2a} at 10%, 50%, and 90% were then plotted from these simulations.

3.4. Metric 2a Discussion (Normal Population Distribution)

The graphs below summarize the general behavior of the metric when the data are drawn from a normal population. Each of the four graphs has a fixed value of “target”. Each different color represents a different “true” value. Solid lines are the 50% quantile and dashed lines are the 10% quantile (the 90% quantile is not shown on these graphs to reduce clutter). See Appendix C for the entire set of graphs generated for Metric 2a for a normally distributed population.

Figure 2: Metric 2a Summary Graphs (Normal Distribution)



The table below shows the median number of samples required to achieve P_{2a} levels of 0.6, 0.7, 0.8, and 0.9, with “target” fixed at 90% lying within the performance limit and “true” varied. Similar tables can be constructed for other choices of “target” and other quantiles of the simulated metric values.

Table 4: Metric 2a Median Number of Samples (Target: 0.9 within spec)

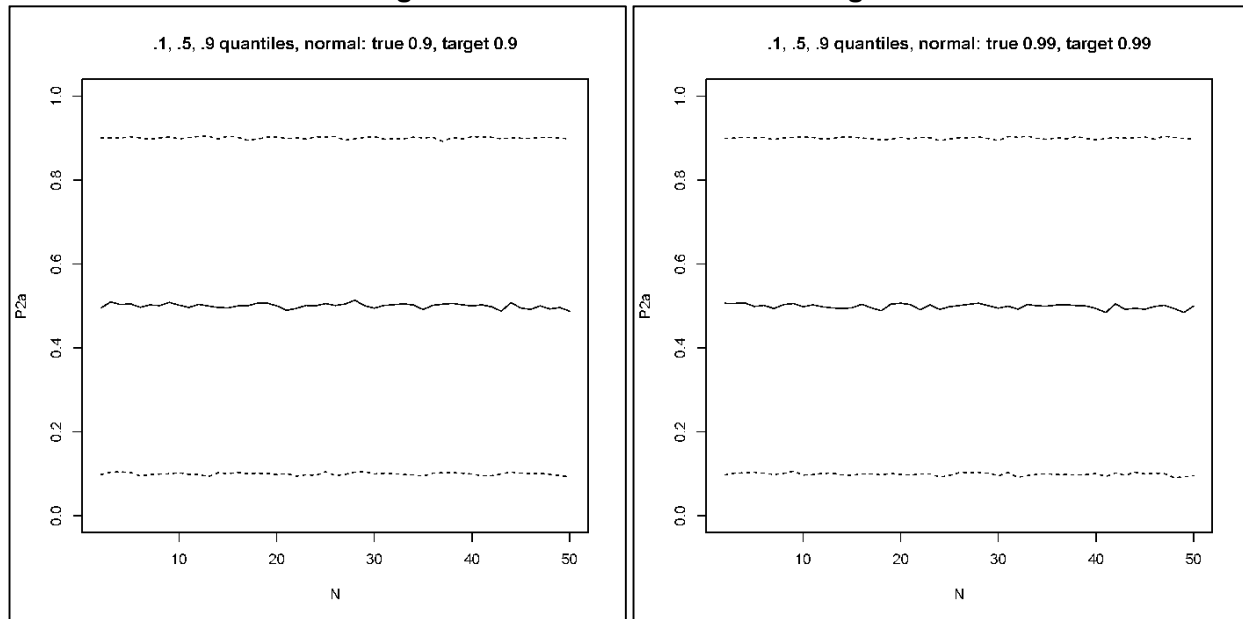
True fraction in spec	P_{2a} level			
	0.6	0.7	0.8	0.9
0.95	3	6	14	30
0.99	2	2	4	7
0.995	2	2	3	5
0.999	2	2	3	4
0.99999	2	2	2	3

General observations include the following:

- Metric 2a is dependent upon the actual conforming proportion (“true”), the proportion that one wishes to protect to (“target”), sampling variation, and N (the number of samples).
- The true conforming proportion depends on characteristics of the population but also on the specification limit used. It is well-known that specification limits are often conservative (or unknown), so the immediate response to a low value of P_{2a} may well be to first examine and revise the spec limit to ensure it is not overly conservative.
- Because of the sampling variation, it is possible that the value of P_{2a} will decrease when more tests are done. This may not be intuitive to users of the metric.
- When $P_{2a} < 0.5$, the implication is that it is more certain that the population does not meet the target.
- There are three distinct regimes of behavior:
 - When “true” \gg “target”, then P_{2a} converges to 1 quickly (*i.e.*, the 10%, 50%, and 90% quantiles converge to 1).
 - When “true” \ll “target”, then P_{2a} converges to 0 quickly (*i.e.*, the 10%, 50%, and 90% quantiles converge to 0).
 - When “true” = “target”, an interesting behavior is seen. The 50% quantile is at a metric value of 0.5, the 10% quantile is at a metric value of 0.1, and the 90% quantile is at a metric value of 0.9. Further, these values are independent of N and of the ratio of “true”/target”. This behavior is illustrated in Figure 3. The metric in this case has a uniform distribution in the range $[0,1]$.
 - The closer “true” and “target” are, the larger the uncertainty is with minimal improvement irrespective of the number of samples tested. This is somewhat disconcerting in that the case of maximum ambiguity is not aided by increasing N . To have a reasonable metric, the target value should be set lower than the true value, and this requires an estimate of the true value.

- For most parameters, it is expected that the “true” conforming proportion will start high by virtue of the design process. In general, margin insufficiency is not common when systems enter the stockpile and the true margin likely corresponds to a conforming proportion of 0.9999 to 0.99999 for a normal population. However the concern is that margin insufficiency may develop over time due to age-related issues. As such, when thinking about selecting a “target” value, it seems to indicate that we don’t want to set “target” so high that it may result in the “true=target” case described above, but we do want to balance that with the objective of being predictive which would suggest higher values of “target”.

Figure 3: Metric 2a when True=Target



- The overly conservative spec limit issue mentioned above also in fact aids us in that it tends to highlight cases before they become performance issues. Given all of this, a “target” value of 0.95 may be a reasonable compromise for most parameters. This is also suggested by the data in Table 5 for “target”=0.95. Note that both the 50% and 10% quantiles achieve relatively high levels of the metric value for values of N around 30, even when “true” is only 0.99.

Table 5: Metric 2a by True (Target: 0.95 within spec)
[10/50/90% quantile values]

		True						
N	Quantile	0.8	0.9	0.95	0.99	0.995	0.9999	0.99999
10	10%	0.00/	0.01/	0.10/	0.51/	0.65/	0.84/	0.98/
	50%	0.02/	0.21/	0.51/	0.86/	0.92/	0.97/	1.00/
	90%	0.33	0.73	0.90	0.99	0.99	1.00	1.00
20	10%	0.00/	0.01/	0.10/	0.69/	0.85/	0.97/	1.00/
	50%	0.00/	0.12/	0.50/	0.94/	0.98/	1.00/	1.00/
	90%	0.08	0.60	0.90	1.00	1.00	1.00	1.00
30	10%	0.00/	0.00/	0.10/	0.81/	0.93/	0.99/	1.00/
	50%	0.00/	0.08/	0.50/	0.98/	0.99/	1.00/	1.00/
	90%	0.02	0.50	0.90	1.00	1.00	1.00	1.00
40	10%	0.00/	0.00/	0.11/	0.88/	0.97/	1.00/	1.00/
	50%	0.00/	0.05/	0.50/	0.99/	1.00/	1.00/	1.00/
	90%	0.00	0.40	0.90	1.00	1.00	1.00	1.00
50	10%	0.00/	0.00/	0.10/	0.93/	0.99/	1.00/	1.00/
	50%	0.00/	0.03/	0.50/	1.00/	1.00/	1.00/	1.00/
	90%	0.00	0.33	0.90	1.00	1.00	1.00	1.00

3.5. Metric 2a Discussion (Lognormal Population Distribution)

One of the questions raised in the sensitivity study was the impact of non-normality on the metric value. While it is possible to test for normality prior to calculating the metrics, it will likely be difficult in many cases to identify non-normality because of small data quantities. Hence a study was done for Metric 2a to see what impact non-normality might have. Two non-normal distributions were chosen to compare to the results for a normal distribution: a lognormal distribution with the same variance as the normal distribution and a lognormal distribution with a variance approximately five times that of the normal distribution.

All distributions were constructed to have the same true proportion lying within the spec limit. This was done for the purpose of simulation by setting the spec limit to the desired true percentile of the standard normal distribution, and then adjusting the lognormal means so that the spec limit also coincided with the same desired true percentile of these distributions. The probability density plots for some example distributions are shown in Figure 4 and 5. These examples illustrate the differences between the normal and lognormal distributions with the same true proportion lying within the spec limit. In general the distributions move farther apart for an increasing true proportion lying within spec limit, and the differences are greater when the lognormal variance is greater than the normal variance. These differences are due to the slow drop-off in probability in the long left tail of the lognormal distribution.

Figure 4: Non-Normal Distributions for Metric 2a: Normal vs. LN Same Variance

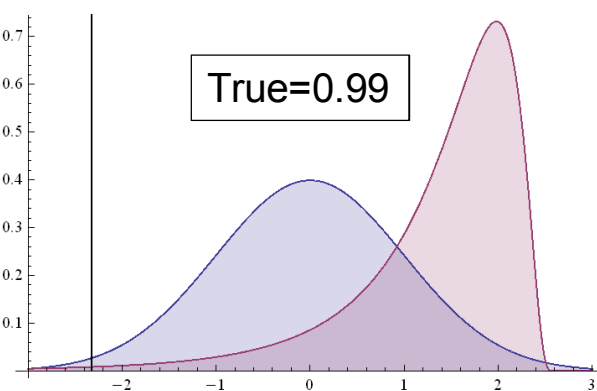
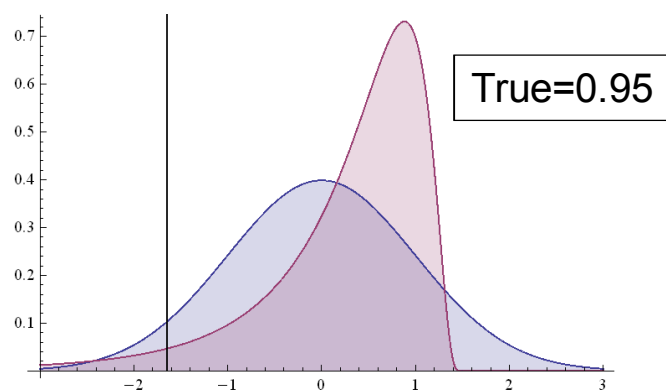
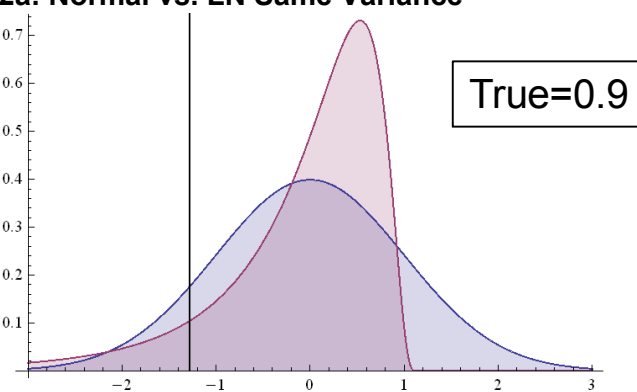
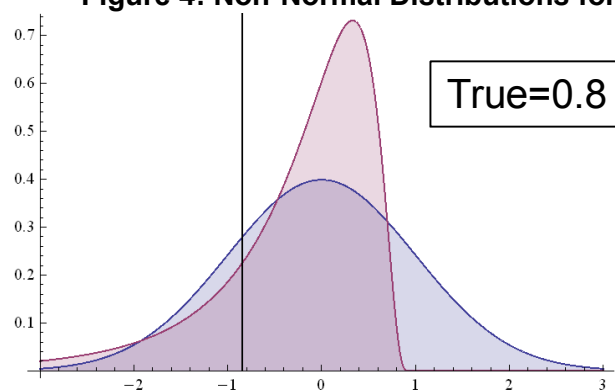
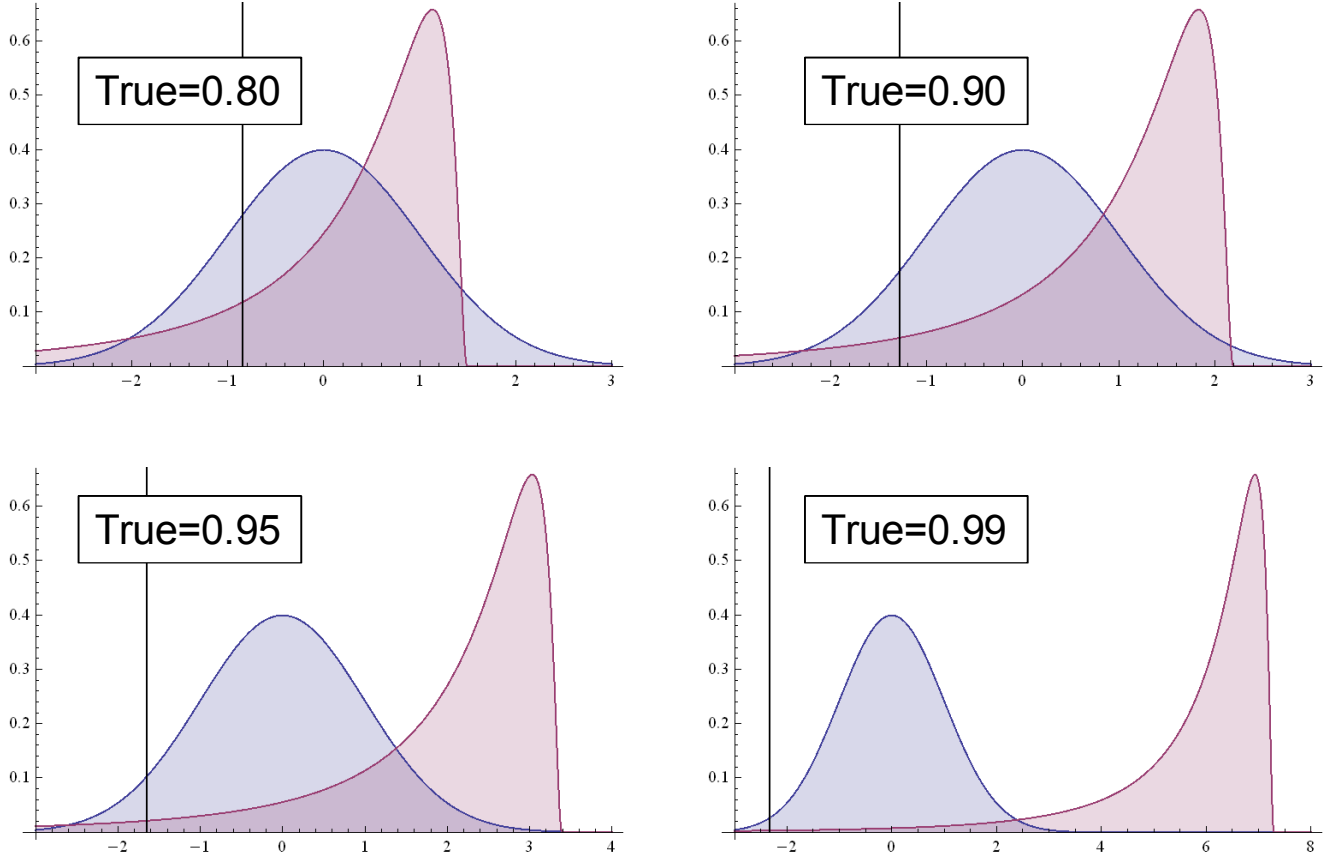


Figure 5: Non-Normal Distributions for Metric 2a: Normal vs. LN Larger Variance



All resulting graphs are shown in Appendix D. Some illustrative examples are shown in Figure 6. Before discussing them in detail, it is helpful to note that the lognormal distributions that we are considering are negatively skewed as shown in Figures 4 and 5. Recall also that the simulations are done using a lower specification limit. It is believed that these account for the different behavior of the metric when comparing the results from the normal vs. lognormal distributions.

Figure 6 (a) shows the typical behavior when $\text{True} \gg \text{Target}$. Figure 6 (b) shows the typical behavior when $\text{True} \ll \text{Target}$. Graphs (c), (d), and (e) show three cases where $\text{True} = \text{Target}$.

In general the results are intuitive given the differences between normal and lognormal distributions seen in Figure 4. The lognormal distribution gives increasingly optimistic metric values as the true proportion lying within spec is increased, since its distribution then diverges further from the normal.

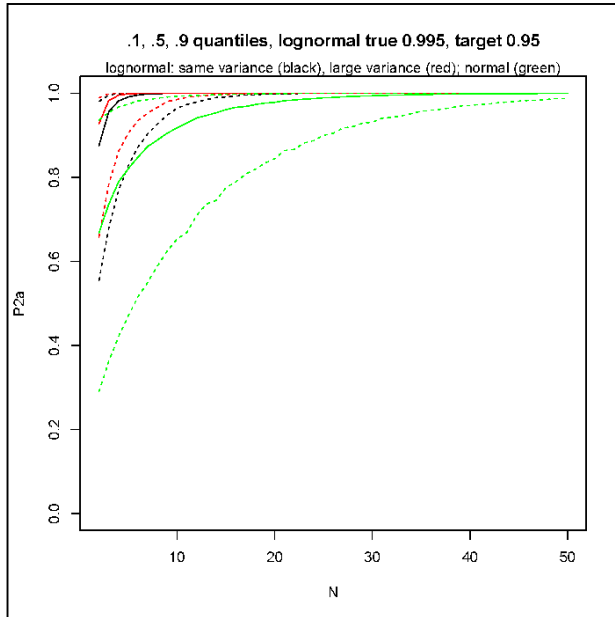
These plots underscore the value of checking for normality prior to calculating the metrics. Clearly it is much more difficult (and misleading) to interpret the metric outcomes if the distribution is skewed. A good follow-up study would be to examine, for example, a bimodal distribution to see the effect of a sub-population on the metric values. It is presumed that this too

could provide misleading results unless a prior normality check was done and the data appropriately subsetting for analysis.

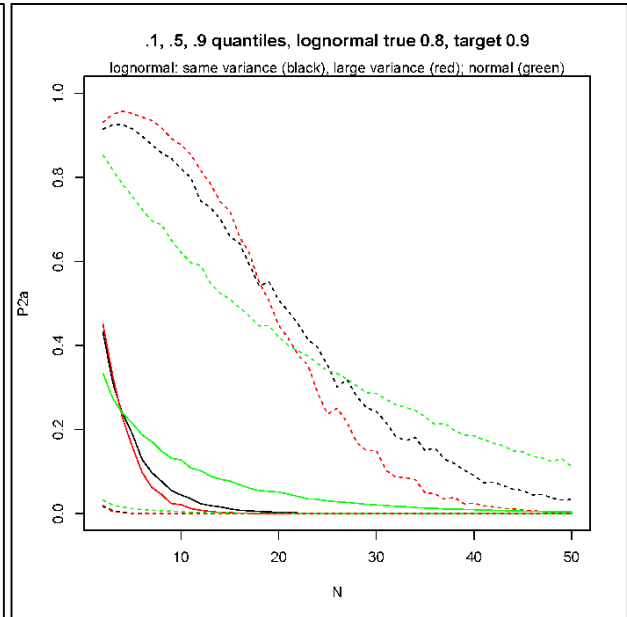
NOTE: On the graphs that follow, the **normal distribution** results are shown in green, the **lognormal with the same variance** in black, and the **lognormal with large variance** in red.

Figure 6: Comparison of Metric 2a for Normal and Lognormal Distributions
[Dashed, solid, dotted lines are .1, .5, .9 quantiles respectively]

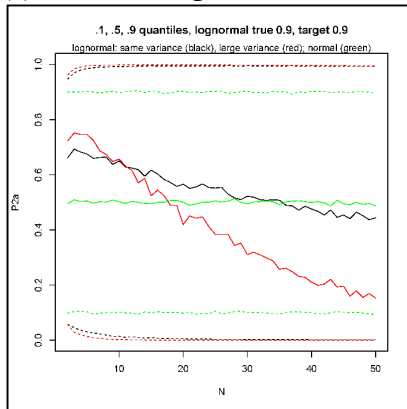
(a) $True \gg Target$



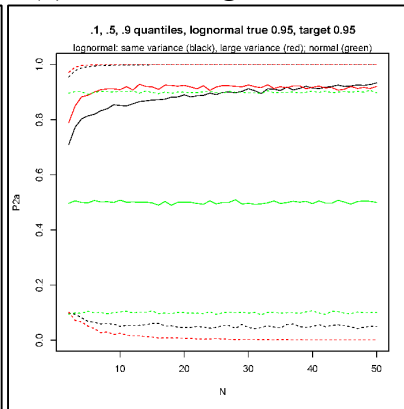
(b) $True \ll Target$



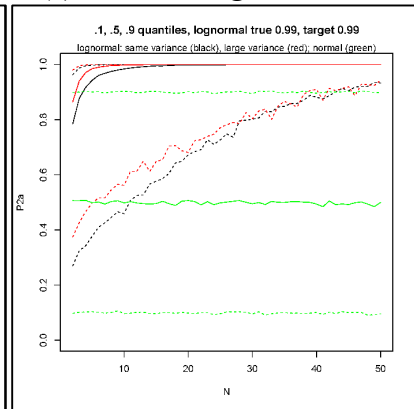
(c) $True = Target = 0.9$



(d) $True = Target = 0.95$



(e) $True = Target = 0.99$



3.6. Metric 2b/2c Discussion

Metrics 2b and 2c are similar to Metric 2a in calculation but represent a population that may be changing over time:

- Metric 2b is the confidence that a proportion of the population does (or does not) exceed a threshold value today, if the data are trended to allow for the possibility that margins may be changing.
- Metric 2c is the confidence that a proportion of the population does (or does not) exceed a threshold value at a specified time in the future, if the data are trended to allow for the possibility that margins may be changing.

Metric 2b/2c use a regression model to predict the property of interest (*e.g.*, from a trend or polynomial fit). One-sided tolerance limits are then calculated for a single, pre-specified time, t , as:

$$Y_{L,t} = \hat{Y}_t - k_t s \quad (11)$$

$$Y_{U,t} = \hat{Y}_t + k_t s \quad (12)$$

Here k_t is the tolerance factor and \hat{Y}_t is the predicted mean response for the specified time, t . Further, s is the estimated standard deviation around the regression line.

The exact analytical expression for the tolerance factor is given by:

$$k_t = T_{v,\delta,\gamma}^{-1} \sqrt{H} \quad (13)$$

in which $T_{v,\delta,\gamma}^{-1}$ is the 100 γ % percentile of the non-central t-distribution with $v = n - p$ degrees of freedom, where p is the number of predictor variables. The non-centrality parameter δ is given by:

$$\delta = \frac{z_P}{\sqrt{H}} \quad (14)$$

in which z_P is the 100 P % percentile of the standard normal distribution, and P is the target conforming proportion of the population that meets the threshold. In addition,

$$H = \sqrt{x_t' (X'X)^{-1} x_t} \quad (15)$$

in which X is the matrix of predictor variables, and x_t is the vector of predictors for the specified time, t .

For the specified time, t , we set $Y_{L,t} = L$ and calculate the confidence:

$$P_{2b/2c} = T_{v,\delta} \left(\frac{\hat{Y}_t - L}{s\sqrt{H}} \right) \quad (16)$$

Similarly, for an upper performance limit:

$$P_{2b/2c} = T_{v,\delta} \left(\frac{L - \hat{Y}_t}{s\sqrt{H}} \right) \quad 17)$$

in which $T_{v,\delta}$ is the cumulative distribution function of the non-central t-distribution with $v = n - p$ degrees of freedom, where p is the number of predictor variables.

Sensitivity studies for Metrics 2b/2c varied sampling and prediction times, as well as the total number of samples:

1. For Metric 2b, the prediction time, t , was set to the end of the sampling period (to represent confidence in margins today).
2. For Metric 2c, prediction time was set to the end of the sampling period plus 15 years.
3. Sampling periods of 10 years and 30 years were evaluated. Results were summarized to compare metric values at $N/30$ samples per year in the 30-year period with metric values at $N/10$ samples per year in the 10-year period.
4. The “target” at prediction time was set to 0.9 proportion conforming.
5. The “true” at prediction time was set to varying levels of 0.95, 0.99, 0.995, 0.999, and 0.9999 proportion conforming.
6. In addition, the true proportion conforming at the start of the sampling period was set to varying levels.

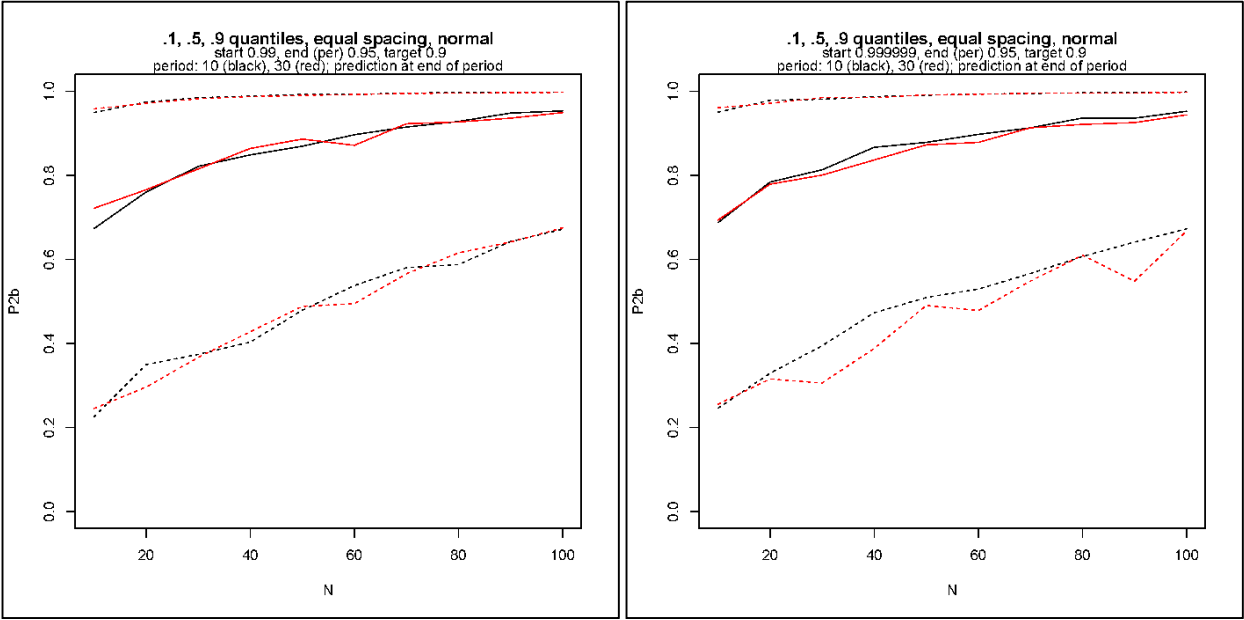
All graphs for Metric 2b are in Appendix E, with graphs for Metric 2c in Appendix F.

General observations for P_{2b} :

- If one examines the graphs for P_{2b} for all conditions, it appears that there is little difference between spreading the samples over a 10-year period or over a 30-year period. Further, it appears that the metric value depends only upon the true conforming at prediction time (*i.e.*, the end of the sampling period, denoted end(per) on the graph) and not upon true conforming at the start of the sampling period (denoted start on the graph). In other words, P_{2b} appears to be independent of the slope.² These two observations are illustrated by data shown on Figure 7.
- P_{2b} is thus mainly dependent upon “end (per)”, target, and N.

² This fact can be verified from Equations 11 through 17 above, which show that Metric 2b/2c does not depend on the starting margin or slope, but only on \hat{Y}_t , the predicted response at time t .

Figure 7: Comparison of Metric 2b For Different Sample Period, Different Start but Same End(per)



• In

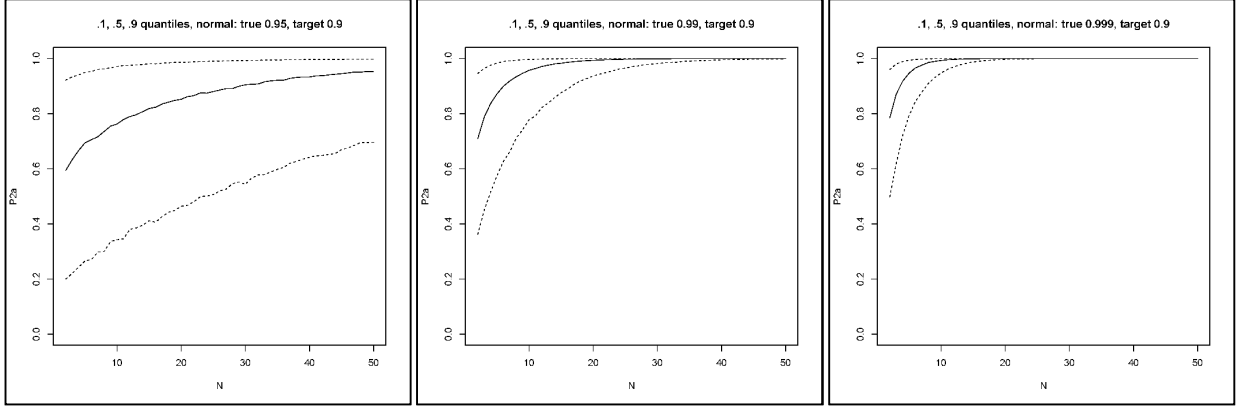
- Figure 8 below, the values for P_{2a} and P_{2b} are compared for similar conditions. Note that the x-axes have different ranges. Graphs (a)-(c) show the results for P_{2a} where the target=0.9 while the true value is varied. This is compared with graphs (c)-(f) for P_{2b} , where again the target=0.9 and the true is varied. These results show that it takes more data to achieve the same confidence when the data are trended by regression analysis to allow for the possibility that margins may be changing.

Figure 8: Comparison of Metric 2a and 2b
[Note that the x-axes have different ranges]

(a) P_{2a} True=0.95 Target=0.9

(b) P_{2a} True=0.99 Target=0.9

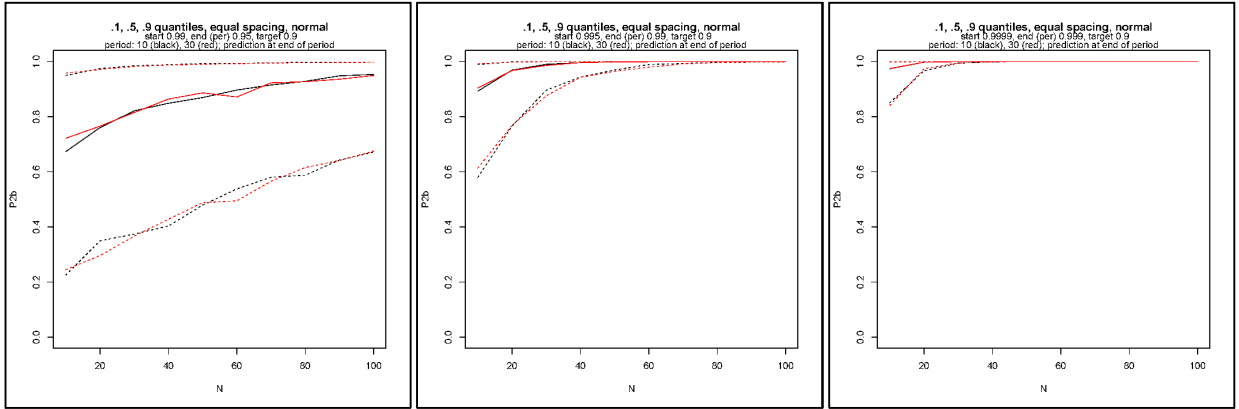
(c) P_{2a} True=0.999 Target=0.9



(d) P_{2b} True(end)=0.95 Target=0.9

(e) P_{2b} True(end)=0.99 Target=0.9

(f) P_{2b} True(end)=0.999 Target=0.9



- The tables below show the median number of samples it took to achieve given metric values for a specified end(per) and target.

Table 6: Metric 2b for End=0.95 (Target: 0.9 within spec)

	P_{2b} level			
Sampling period	0.6	0.7	0.8	0.9
30 years	10	20	30	60
10 years	10	20	40	70

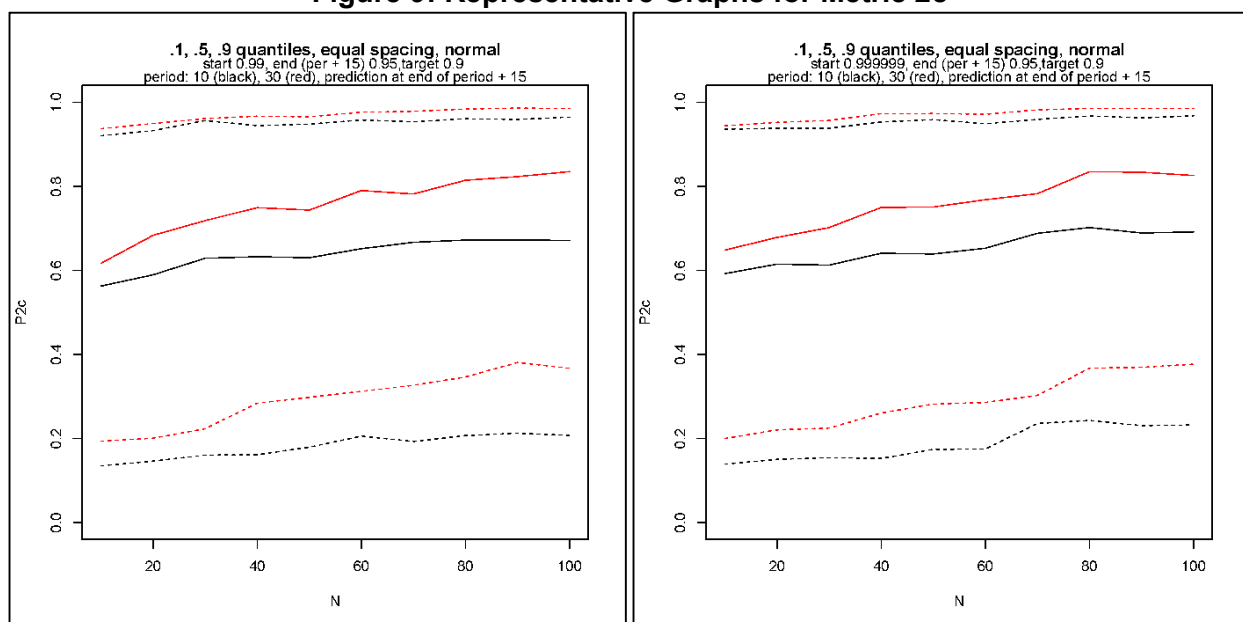
Table 7: Metric 2b for End=0.99 (Target: 0.9 within spec)

	P_{2b} level			
Sampling period	0.6	0.7	0.8	0.9
30 years	10	10	10	20
10 years	10	10	10	20

General observations for P_{2c} are as follows:

- As with P_{2b} , there is no dependence upon the “start” value.
- Compared to P_{2b} , for P_{2c} there is a greater difference between spreading the sample over a 10-year period vs. over a 30-year period. This difference is readily seen by comparing representative graphs for P_{2b} (Figure 7) and P_{2c} (Figure 9).

Figure 9: Representative Graphs for Metric 2c



- From Equations 8 through 10, it can be seen that the systematic (as opposed to random) differences between the 10-year and 30-year sampling periods are due to the matrix H . Table 8 below presents the results of some additional calculations that were done to explore the behavior of the matrix H for various sampling and prediction periods.

**Table 8: Matrix H and Tolerance Factor k
(Confidence=0.9, Target=0.9)**

Sampling period	Sample rate	Prediction year	H	k
Prediction at end of period (Metric 2b)				
30 years	1/yr	30	0.127	1.85
15 years	2/yr	15	0.121	1.84
10 years	3/yr	10	0.115	1.83
6 years	5/yr	6	0.105	1.81
Prediction at end plus 15 years (Metric 2c)				
30 years	1/yr	45	0.421	2.21
15 years	2/yr	30	0.898	2.60
10 years	3/yr	25	1.570	2.99
6 years	5/yr	21	3.533	3.81

- The effect of sampling period and prediction time is further illustrated by calculating tolerance limits for simulated data sets at 3/yr over 10 years (Figure 10) and 1/yr over 30 years (Figure 11). Tolerance limits for 90 percent confidence and 90 percent coverage are shown, for simulated data from a population whose true conforming proportion trends down from 0.9999 to 0.99 over the sampling period plus 15 years. The interpretation of Figures 10 and 11 is that the confidence that 90% of units exceed zero (which was the spec limit used for simulation) is exactly 90 percent when the lower tolerance band crosses the x-axis, and less than 90 percent for later times. Comparison of some randomly selected examples in Figures 10 and 11 shows that the prediction uncertainty is clearly magnified in the 10-year case, due to extrapolation with a less precise estimate for slope.

Figure 10: Tolerance Limits for Simulated Data Sets at 3/yr Over 10-Years
 (confidence = 0.9, target = 0.9, true trends from 0.9999 to 0.99 over 25 years)

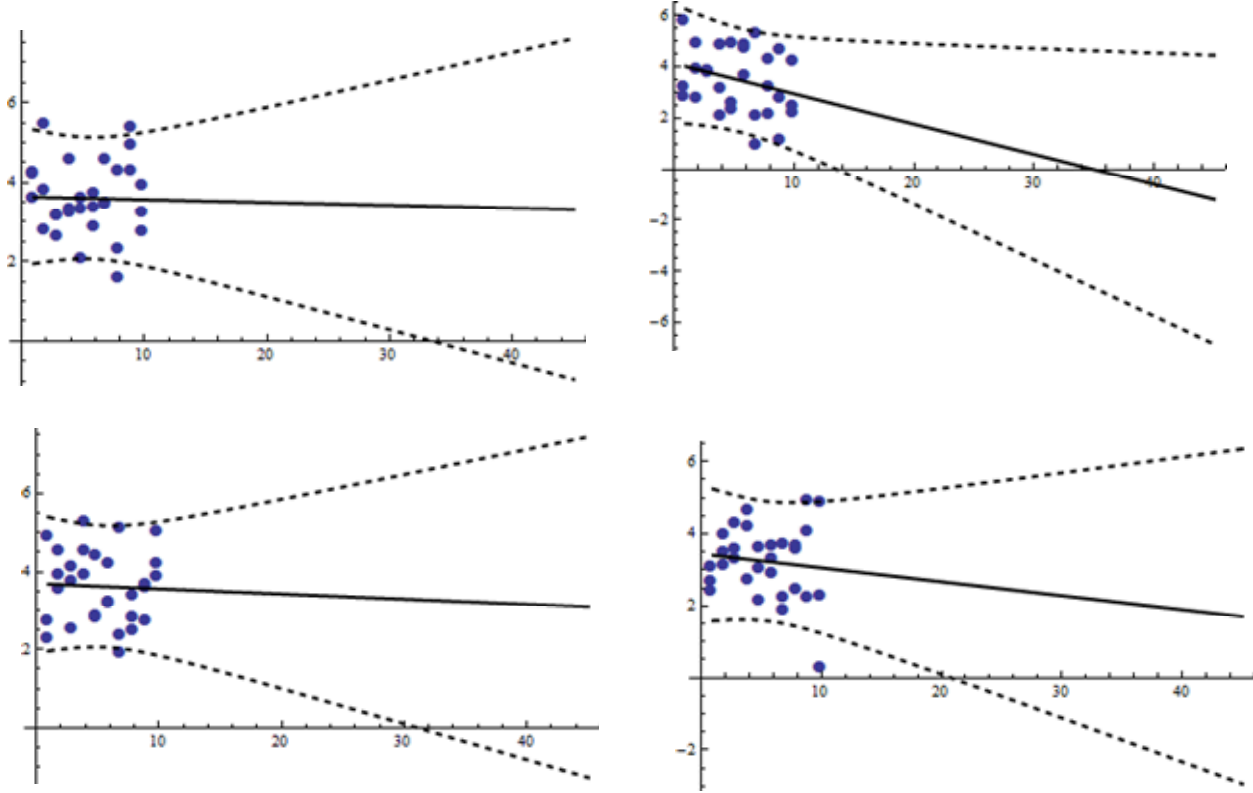
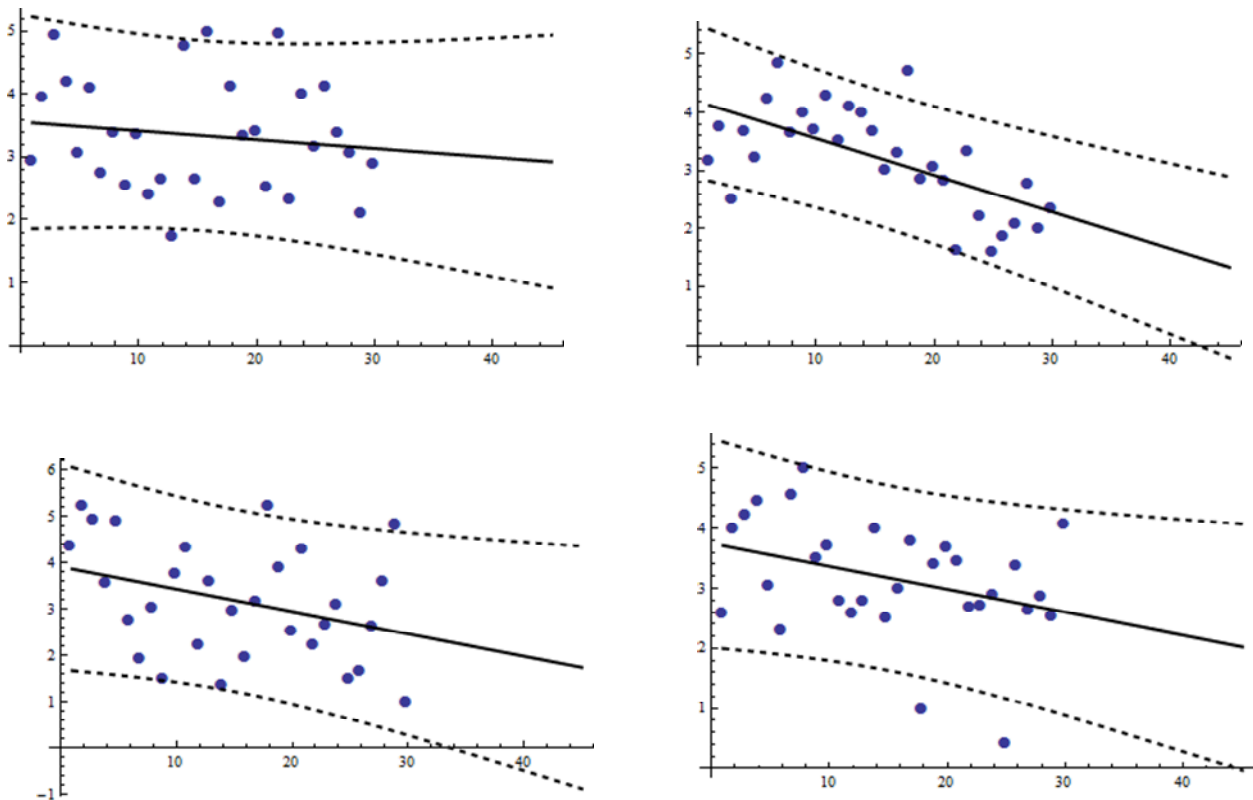


Figure 11: Tolerance Limits for Four Simulated Data Sets at 1/yr Over 30-Years
(confidence = 0.9, target = 0.9, true trends from 0.9999 to 0.99 over 45 years)



- It is apparent from
- Table 8 (above) that Metric 2b (representing the ability to estimate current conditions) is only slightly improved when the same number of samples are taken over a shorter period that is closer to the current time (H and k values for Metric 2b in the top half of the table are relatively constant compared to H and k values for Metric 2c in the bottom half of the table). On the other hand, Metric 2c (representing the ability to project future conditions) is greatly improved by taking samples over an extended period to better estimate slope. This suggests that a sustained sampling program over time would be a better allocation of resources, relative to a short burst of sampling activity.
- Note that Metric 2c was worse for the 10-year than 30-year sampling periods when prediction times were Year 25 and Year 45, respectively. The 10-year sampling period will clearly perform worse still if a Year 45 prediction time is used for both sample periods.
- As with Metric 2a, the immediate response to a low value of Metric 2b/2c may well be to first examine and revise the spec limit to ensure it is not overly conservative.
- Another response may be to get more current data to better estimate the slope and current/future values.
- Measurement error is another possible cause of noisy data. Measurement error will tend to lower the calculated metric, since the tolerance limit approach assumes that all of the

observed scatter in the data is due to true unit-to-unit variation. Various steps could reduce the effect of measurement error, *e.g.*:

- estimate the contribution of measurement error and revise the metric to account for it.
- develop a less noisy measurement procedure.
- Another response is to simply collect additional data. However, although getting more data will improve the precision of the slope estimate, it will not improve the estimated scatter around the regression line. Hence measurement error will still lead to conservative, under-estimates of the metric.

3.7. Metric 3a/3b Discussion

The purpose of this metric is to assess data adequacy to detect departures from previous predictions from a science model. More specifically, this metric is the probability that a surveillance sample will reveal a specified change in the mean or standard deviation of its measured parameter from that previously predicted by science models. Metric 3 addresses on-going validation of the predictive capability of science models. The time period during which this metric is applied is determined from guidance of subject matter experts about the need for on-going validation. The SMEs may take into account subjective information about model quality in specifying the time period during which the on-going validation is conducted.

The power for model validation is addressed by two sub-metrics:

P_{3a} : The power to detect a 2-sigma shift in mean from the model prediction. This calculation also uses the model prediction of the population standard deviation, σ .

P_{3b} : The power to detect a doubling in standard deviation from the model prediction.

The test statistic for a difference in mean from a science-based model is the chi-square score

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\sigma_i^2} \quad 18)$$

in which \hat{Y}_i is the science-based model prediction corresponding to the i^{th} measurement. In general the science-based model includes a Gaussian model of unit-to-unit variation in which σ_i is the model prediction for the standard deviation among units.

Under the null hypothesis, the sample measurements are drawn from distributions predicted by the model, so that

$$Y_i \sim N(\mu_i = \hat{Y}_i, \sigma_i^2) \quad 19)$$

and the chi-square score follows the chi-square distribution with n degrees of freedom.

We propose to reject the null hypothesis when the probability of the chi-square score is less than 0.05. This gives the criterion for rejection of the null hypothesis

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\sigma_i^2} > X_{n,0.95}^{-1} \quad (20)$$

in which the critical value, $X_{n,0.95}^{-1}$, is the 95th percentile of the chi-square distribution with n degrees of freedom.

If the actual measurement means differ from the model predictions so that $\mu_i = \hat{Y}_i + d_i$, the chi-square score follows a non-central chi-square distribution with n degrees of freedom and non-centrality parameter

$$\lambda = \sum_{i=1}^n \left(\frac{d_i}{\sigma_i} \right)^2 \quad (21)$$

Using our five-percent significance test, the power to detect a set of shifts, d_i , is

$$\text{Power} = P_{3a} = 1 - X_{n,\lambda}(X_{n,0.95}^{-1}) \quad (22)$$

in which $X_{n,\lambda}$ is the cumulative distribution function of the non-central chi-square distribution with n degrees of freedom and non-centrality parameter λ .

P3b: Power to detect a difference in standard deviation from a science-based model

The test statistic for a difference in standard deviation from a science-based model is the chi-square score

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\sigma^2} \quad (23)$$

in which \hat{Y}_i is the science-based model prediction corresponding to the i^{th} measurement. In general the science-based model includes a Gaussian model of unit-to-unit variation in which σ is the model prediction for the standard deviation among units.

Under the null hypothesis, the sample measurements are drawn from distributions predicted by the model, so that

$$Y_i \sim N(\mu_i = \hat{Y}_i, \sigma^2) \quad (24)$$

and the chi-square score follows the chi-square distribution with n degrees of freedom. We propose to reject the null hypothesis when the probability of the chi-square score is less than 0.05. This gives the criterion for rejection of the null hypothesis

$$\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{\sigma^2} > X_{n,0.95}^{-1} \quad (25)$$

in which the critical value, $X_{n,0.95}^{-1}$, is the 95th percentile of the chi-square distribution with n degrees of freedom.

If the actual standard deviation, σ_i , is some multiple such that $\sigma_i = m\sigma$ the chi-square score divided by m^2 follows the chi-square distribution. Using our five-percent test of significance, the power to detect the difference in standard deviation is therefore

$$\text{Power} = P_{3b} = 1 - X\left(\frac{X_{n,0.95}^{-1}}{m^2}\right). \quad 26)$$

Similar to Metric 1a and 1b, Metrics 3a and 3b will be discussed together because of their similarity. These metrics have the very attractive feature that, as defined, their values are a function only of N. The sensitivity study graphs are shown in Figure 12 with the information for Metric 3a and Metric 3b summarized in Table 9 and Table 10 respectively.

General observations include the following:

- These metrics are formulated to be independent of the actual mean and standard deviation (SD) values for the population as well as the specification limit.
- As noted in the graphs, there is no direct dependence upon the time window in which one is counting tests, but that will clearly affect the value of N.

Figure 12: Metric 3a/3b Summary Graphs

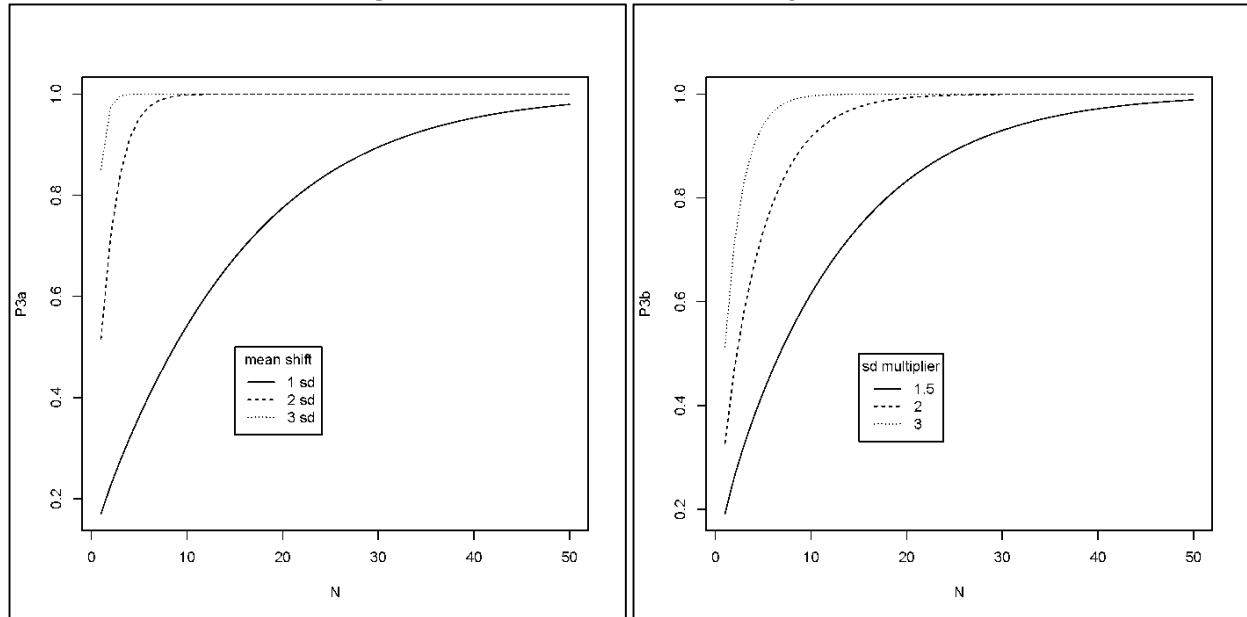


Table 9: Metric 3a by Size of Shift in Mean (in Standard Deviations, SD)

	P_{3a} level			
Mean shift size	0.6	0.7	0.8	0.9
1 SD	12	17	22	31
2 SD	2	2	3	4
3 SD	1	1	1	2

Table 10: Metric 3b by Size of Increase in Standard Deviation (in Standard Deviations, SD)

	P_{3b} level			
SD increase	0.6	0.7	0.8	0.9
1.5 x	10	14	18	26
2 x	4	5	7	10
3 x	2	2	3	4

- There is a major difference in sample quantities required to achieve the same level of P_{3a} between looking for a one standard deviation (SD) vs. a two SD shift in mean. Given that (1) these are limited to quantities of units tested at the upper end of the age range and (2) that these tests must be done on an on-going basis, it is unlikely that the sample rates needed to detect a one SD shift are sustainable.
- This is likewise true for P_{3b} . Even a 2x increase in SD requires a relatively large number of samples on an on-going basis, given that data lose their currency.
- Note that this calculation presumes a sudden shift in mean and/or standard deviation. Analysis must be done on the test data itself to determine if in fact there has been a statistically significant change in the variable or just in the measurement system.
- There are at least two considerations for selecting the time window for Metrics 3a/3b. For the most part, the time period for which the data are included is determined from guidance of subject matter experts about the need for on-going validation. The SMEs may take into account subjective information about model quality in specifying the time period during which the on-going validation is conducted. However it is likely prudent to have an upper bound for this time period (10 years, perhaps) to cap the risk at some point. It is certainly possible to place too much confidence in a model and position ourselves to miss the opportunity to detect something unexpected.
- The size of the mean shift or standard deviation increase can be set at the same number of standard deviations for all parameters. However it is probably also reasonable to consider the option of letting this vary from parameter to parameter based upon that parameter's margin; i.e., parameters with a larger margin could be set to detect a larger mean or standard deviation shift than those with smaller margin. If this approach is used, the rationale for selecting the shift to be detected should be documented.

3.8. Metric 4 Discussion (Linear Trend)

The purpose of Metric 4 is to quantify the power to detect a trend that would eliminate margin over a period of 15 years from the present. The power to detect a trend is particularly important when data trend analysis is used as the primary means to identify changes of concern for further study.

Metric 4 assumes that analysts test for a trend by fitting the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad 27)$$

where x_i , $i = 1, \dots, n$ are the unit ages at test, α and β are unknown, and ε_i are independent normally distributed random errors with expected value 0 and unknown variance σ^2 , and Y_i , $i = 1, \dots, n$ are observed variables measurements.

The test statistic to test for a significant trend is the t-score:

$$\frac{\hat{\beta}}{s_{\hat{\beta}}} \quad 28)$$

in which $\hat{\beta}$ is the ordinary least squares estimate of the slope, β , and $s_{\hat{\beta}}$ is the standard deviation of $\hat{\beta}$:

$$s_{\hat{\beta}} = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad 29)$$

Further, s is the estimated standard deviation around the regression line:

$$s = \sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n - 2}} \quad 30)$$

Under the null hypothesis that the slope β is equal to 0, the t-score follows the t-distribution with $\nu = n - 2$ degrees of freedom. Therefore, the probability of observing a t-score as large or larger (in absolute value) is for a negative trend

$$T_{\nu} \left(\frac{\hat{\beta}}{s_{\hat{\beta}}} \right) \quad 31)$$

and for a positive trend

$$1 - T_{\nu} \left(\frac{\hat{\beta}}{s_{\hat{\beta}}} \right) \quad 32)$$

in which T_{ν} is the cumulative distribution function of the t-distribution with $n-2$ degrees of freedom.

We propose to reject the null hypothesis when the probability of the t-score is less than 0.05. This gives the criterion for rejection of the null hypothesis

$$\left| \frac{\hat{\beta}}{s_{\hat{\beta}}} \right| > T_{\nu, 0.95}^{-1} \quad 33)$$

in which the critical value, $T_{v,0.95}^{-1}$, is the 95th percentile of the t-distribution with $v = n - 2$ degrees of freedom.

If the actual trend is some value $\beta_0 \neq 0$, the t-score follows a non-central t-distribution with $v = n - 2$ degrees of freedom and non-centrality parameter

$$\delta = \frac{\beta_0 \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sigma} \quad (34)$$

Using our five-percent test of significance, the power to detect a trend is therefore calculated from the non-central t-distribution

$$\text{Power} = 1 - T_{v,\delta}(T_{v,0.95}^{-1}) \quad (35)$$

in which $T_{v,\delta}$ is the cumulative distribution function of the non-central t-distribution with $v = n - 2$ degrees of freedom and non-centrality parameter δ .

The non-centrality parameter and therefore the power calculation depends on the population standard deviation, σ , which in general is not known. As a point estimate, we propose to estimate δ by substituting the sample standard deviation, s , for σ :

$$\hat{\delta} = \frac{\beta_0 \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{s} = \frac{\beta_0}{s_{\hat{\beta}}} \quad (36)$$

In addition, a lower bound can be calculated for the non-centrality parameter using an exact calculation presented by Taylor and Muller (1995). The lower bound for the non-centrality parameter derives from the distributional relation:

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2 \quad (37)$$

Hence, e.g.

$$P \left[\frac{(n-2)s^2}{\sigma^2} < X_{n-2,0.05}^{-1} \right] = 0.05 \quad (38)$$

in which $X_{n-2,0.05}^{-1}$ is the 5th percentile of the chi-square distribution with $n-2$ degrees of freedom. Then

$$P \left[\delta^2 = \frac{\beta_0^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} < \frac{\beta_0^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(n-2)s^2} X_{n-2,0.05}^{-1} \right] = 0.05 \quad (39)$$

So a 95-percent one-sided lower bound estimate for the non-centrality parameter, δ , is

$$\hat{\delta}_{0.05} = \sqrt{\frac{\beta_0^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(n-2)s^2}} X_{n-2,0.05}^{-1} = \frac{\beta_0}{s_{\hat{\beta}}} \sqrt{\frac{X_{n-2,0.05}^{-1}}{n-2}} \quad (40)$$

Similarly a 95-percent one-sided upper bound estimate for the non-centrality parameter, δ , is

$$\hat{\delta}_{0.95} = \sqrt{\frac{\beta_0^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(n-2)s^2}} X_{n-2,0.95}^{-1} = \frac{\beta_0}{s_{\hat{\beta}}} \sqrt{\frac{X_{n-2,0.95}^{-1}}{n-2}} \quad (41)$$

Because of the strict monotone dependence of the non-central t-distribution on noncentrality, the lower and upper bounds for power can be found from

$$\text{Power}_{0.05} = 1 - T_{\nu, \hat{\delta}_{0.05}}(T_{\nu,0.95}^{-1}) \quad (42)$$

$$\text{Power}_{0.95} = 1 - T_{\nu, \hat{\delta}_{0.95}}(T_{\nu,0.95}^{-1}) \quad (43)$$

We propose to “grade” data adequacy in terms of the power to detect a trend that eliminates margin in the weakest 10% of units over a period of 15 years. The critical trend is defined as:

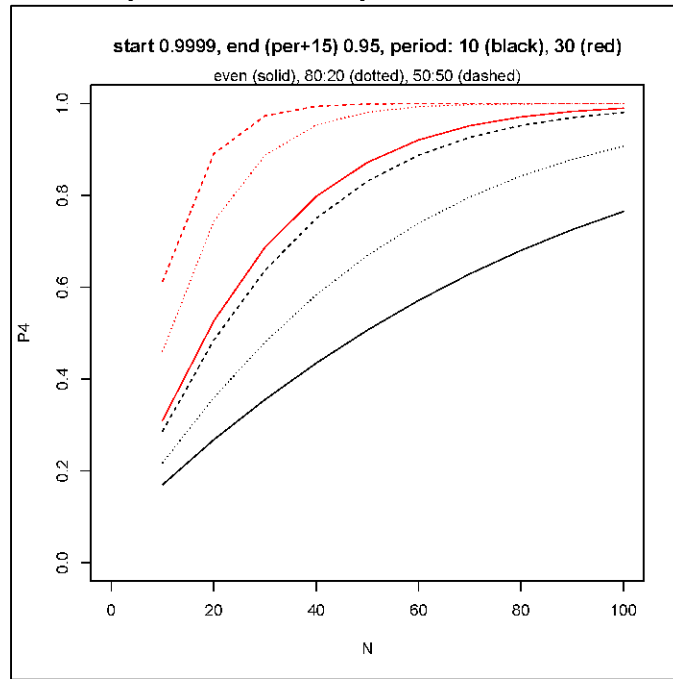
$$\beta_{crit} = \frac{(Y_0 - L) - 1.28\sigma}{15} \quad (44)$$

in which $(Y_0 - L)$ is the initial margin and σ is the standard deviation among units. Note that, at the extreme, if the initial margin is only 1.28σ then a slope of zero would be critical.

The simulation study examined the behavior of Metric 4 under a range of temporal behaviors including linear and various non-linear trends.

Graphs for the simulation study of Metric 4 for the linear trend case are in Appendix H. Figure 13 shows a graph for a single case, and Table 11 and Table 12 are the same case except for the period of time over which the samples are distributed.

Figure 13: Representative Graph for Metric 4, Linear Trend



General observations include the following:

- In practice, Metric 4 will be a random variable due to the need to estimate the initial mean value, Y_0 , and population standard deviation, σ . For the purpose of the simulation study, the behavior of Metric 4 was examined by treating Y_0 and σ as known. The results of the simulation study correspond to typical (average) values for Metric 4.
- A review of all of the graphs indicates that spreading the samples over 30 years results in a higher metric value than spreading them over 10 years. This is consistent with the improved precision of the slope estimate that was also noted for Metric 2b/2c.
- Distribution of the samples also has a large effect on the metric value. Not surprisingly, the 50-50 distribution (half of the samples taken at the starting time and half at the ending time) is more effective for identifying the behavior of interest in this case, a linear trend. 80-20 is the next best, and evenly spaced samples yields the lowest metric value. It is important to not read too much into this, though. While 50-50 may be the optimal distribution for this particular behavior, it will be less effective at finding other types of defectiveness. This was explored to some extent in a sensitivity study using non-linear trends (described in the next section).

- The power to detect a trend depends on both the slope and the standard deviation as combined into the signal-to-noise ratio (SNR):

$$SNR = \frac{\mu_{start} - \mu_{end}}{\sigma} \quad 45)$$

Note that the SNR is equal to the starting and ending quality indices:

$$SNR = \frac{\mu_{start} - L}{\sigma} - \frac{\mu_{end} - L}{\sigma} \quad 46)$$

The quality indices determine the start and end proportion lying within the spec limit, which is the factor shown on the graphs.

- For cases where the SNR is small (e.g., start = 0.95, end = 0.9), the metric value will be low even for relatively large N .

Table 11: Metric 4, Sampling over 30 Years (Start=0.9999, Start+15 Years=0.95)

	P4 level			
Sample distn.	0.6	0.7	0.8	0.9
Even	30	40	50	60
80:20	20	20	30	40
50:50	10	20	20	30

Table 12: Metric 4, Sampling over 10 Years (Start=0.9999, Start+15 Years=0.95)

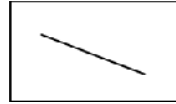
	P4 level			
Sample distn.	0.6	0.7	0.8	0.9
Even	70	90	120	160
80:20	50	60	80	100
50:50	30	40	50	70

3.9. Metric 4 Discussion (Non-Linear Trend)

The following cases for trend scenarios were developed. These profiles were based upon past experiences with real data. The intent was to try to study different trend types with variations in their characteristics:

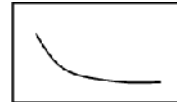
A. Linear

- Vary slope, constant variance
- Vary slope, vary variance



B. Asymptote

- Vary rate of decay, constant variance
- Vary rate of decay, vary variance



C. Step function

- Vary time period over which degradation occurs
- Vary delta in means, constant variance
- Vary delta in means, vary variance



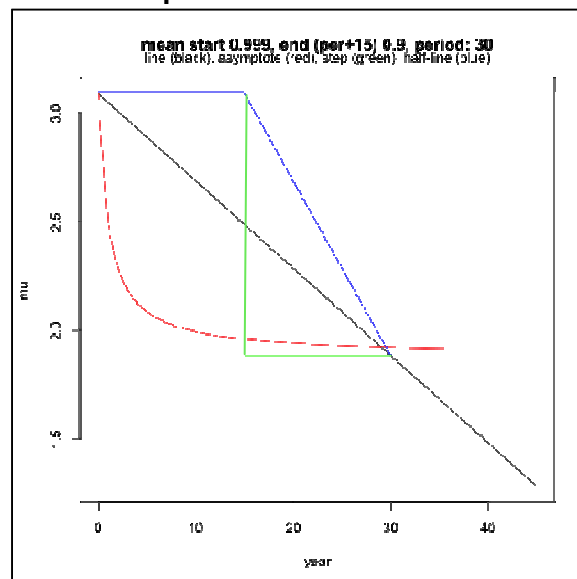
D. Half-line

- Vary time period over which degradation occurs
- Vary delta in means, constant variance
- Vary delta in means, vary variance



However different trend scenarios were studied only to a limited extent and for only one of the metrics, P4. For each of the four trend curves, the beginning and ending points were set to be the same. For the step function and half-line, the transition occurred at the mid-point in time. This is shown in Figure 14.

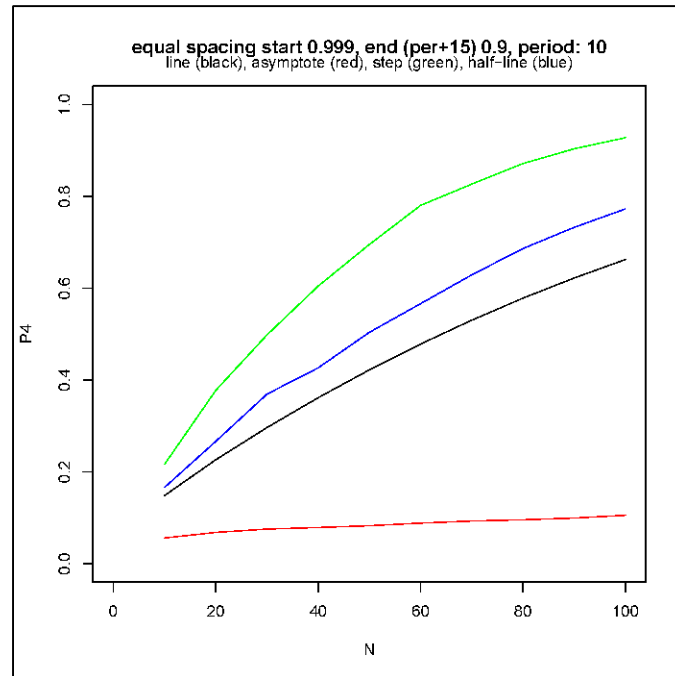
Figure 14: Example of Linear and Non-Linear Trends



Studies that explored the effect of non-linear trends on the Metric 4 calculation are included in Appendix I.

Figure 15 shows a representative example from these calculations, where the different colors indicate the difference trend assumptions.

Figure 15: Representative Graph for Metric 4, Comparison of Linear and Non-Linear Trends



General observations include the following:

- Only the equal spacing sampling cases are displayed. The 50:50 and 80:20 sampling schemes yielded exactly the same metric results for each trend line. This is due to the fact that the problem was set up such that the beginning points and end points were chosen to be identical for each trend line. Since those two sampling schemes only sample at the beginning and ending, it doesn't matter what happened in between vis-à-vis the metric calculation. This actually raises a good point about P4. While the 50:50 sampling strategy is the best for detecting a linear trend, clearly there is no information gained about performance between the points in time defined as the beginning and ending. If one wants to be responsive to behaviors other than linear trends, 50:50 is not necessarily the best approach.
- This reinforces that one should always review the raw data on an on-going basis rather than just relying on the metrics to highlight issues.
- Spreading the samples over 30 years always yields a higher P4 value than spreading them over 10 years.

- The asymptote was the worst case for all variants of starting margin and period. This is likely because there is so little data obtained prior to the large decrease in value, which happens quite early in life.
- The half-line and step function have higher metric values than the linear trend. This corresponds with one's intuition that the slope change after the mid-point would appear to be more dramatic than that for the linear trend.

4. CONCLUSIONS AND RECOMMENDATIONS

This sensitivity study of the proposed set of statistical metrics was performed to evaluate the behavior of the metrics within a range of parameter values that need to be specified to fully define the metrics. The parameters are related to risk. What types of defectiveness or magnitude of problems do surveillance managers want to have a high confidence of detecting? Knowing these risk parameters allows the determination of a sample size to achieve the stated level of risk.

The study allowed us to determine that the following could serve as reasonable values for metric calculation:

- 10% defectiveness level;
- 10% level for fraction of population with low margin;
- 4 year look-back for data currency;
- 15 year prediction interval;
- 2-sigma shift in mean for model invalidation;
- Doubling of standard deviation for model invalidation.

These results can be used in concert with other technical and programmatic factors to make a final decision on appropriate risk levels.

The team concluded that all metrics evaluated showed reasonable behavior with the recommended risk parameters. However the team noted that Metric 4, whose purpose is to convey the confidence in detecting trends, demonstrated that surveillance data may often not be adequate to rule out a critical trend with high confidence, especially in low margin situations. For this reason, Metric 4 was not recommended to be deployed further as a surveillance metric. Instead the team recommended that the Metrics 3b and 3c be routinely deployed to calculate the confidence in current and future margins when allowing for a possible trend, unless there is high confidence that no trend is present.

APPENDIX A: R CODE

Metric 1a

```
#metric Pla

p.v=c(.01, .05, .10, .25) #problem sizes 1% etc

pop.v=c(100,500) #population sizes

n1=seq(1,50,1) #sample sizes
n2=seq(1,250,1) #sample sizes


#infinite size population
n1.len=length(n1)
n1.inf.ans=NULL

y.inf=NULL
for(i in 1:4){
p=p.v[i]
m1.inf=1-(1-p)^n1
n1.inf.ans=rbind(n1.inf.ans,m1.inf)
y.inf=rbind(y.inf,m1.inf)
}


# treat as finite size population, population 500
n.pop=pop.v[2]
n1.ans=NULL

y=NULL
for(i in 1:4){
p=p.v[i]
n.p=n.pop*p
m1.small=1-dhyper(n1,n.pop-n.p,n.p,n1)
n1.ans=rbind(n1.ans,m1.small)
y=rbind(y,m1.small)
}


# treat as finite size population, population 100
n.pop=pop.v[1]
n1b.ans=NULL

y.b=NULL
for(i in 1:4){
p=p.v[i]
n.p=n.pop*p
m1.small=1-dhyper(n1,n.pop-n.p,n.p,n1)
n1b.ans=rbind(n1b.ans,m1.small)
y.b=rbind(y.b,m1.small)
}
```

```

pdf("metric Pla plots.pdf")
plot(x,y,type="n",xlab="N",ylab="Pla")
for(i in 1:4){
  lines(n1,n1.ans[i,],lty=i)
}
for(i in 1:4){
  lines(n1,n1b.ans[i,],lty=i,col="red")
}
legend(30,.23,c(".01",".05",".10",".25"),lty=c(1,2,3,4),title="defect
size")
title("100 (black) and 500 (red) population, hypergeometric")

#minimum sample size for Pla> prob level, population 500, hypergeometric
counts.p=c(.6,.7,.8,.9) #prob level
counts=matrix(rep(0,4*4),ncol=4)

for(i in 1:4){
  for(j in 1:4){
    counts[i,j]= min(n1[n1.ans[i,]>=counts.p[j]])
  }
}
counts

#minimum sample size for Pla> prob level, population 100, hypergeometric
counts.p=c(.6,.7,.8,.9) #prob level
counts.1b=matrix(rep(0,4*4),ncol=4)

for(i in 1:4){
  for(j in 1:4){
    counts.1b[i,j]= min(n1[n1b.ans[i,]>=counts.p[j]])
  }
}
counts.1b

# infinite size population
n2.inf.ans=NULL

y2.inf=NULL
for(i in 1:4){
  p=p.v[i]
  m1.inf=1-(1-p)^n2
  n2.inf.ans=rbind(n2.inf.ans,m1.inf)
  y2.inf=rbind(y2.inf,m1.inf)
}

# population 500, treat as finite size population
n.pop=pop.v[2]
n2.ans=NULL

y2=NULL
for(i in 1:4){
  p=p.v[i]
  n.p=n.pop*p

```



```

m1.small=1-dhyper(n2,n.pop-n.p,n.p,n2)
n2.ans=rbind(n2.ans,m1.small)
y2=rbind(y2,m1.small)
}

x2=rep(n2,4)
plot(x2,y2,type="n",xlab="N",ylab="Pla")
for(i in 1:4){
lines(n2,n2.ans[i,],lty=i)
}
for(i in 1:4){
lines(n2,n2.inf.ans[i,],lty=1,col="red")
}
legend(50,.23,c(".01",".05",".10",".25"),lty=c(1,2,3,4),title="defect
size")
title("hypergeometric (black) for 500 population, binomial (red)")

#
counts.p=c(.6,.7,.8,.9)
counts.2=matrix(rep(0,4*4),ncol=4)

for(i in 1:4){
for(j in 1:4){
counts.2[i,j]= min(n2[n2.ans[i,]>=counts.p[j]])
}
}
counts.2

counts.p=c(.6,.7,.8,.9)
counts.2inf=matrix(rep(0,4*4),ncol=4)

for(i in 1:4){
for(j in 1:4){
counts.2inf[i,j]= min(n2[n2.inf.ans[i,]>=counts.p[j]])
}
}
counts.2inf

# infinite size population
n3=seq(1,500,1)
n3.inf.ans=NULL

y3.inf=NULL
for(i in 1:4){
p=p.v[i]
m1.inf=1-(1-p)^n3
n3.inf.ans=rbind(n3.inf.ans,m1.inf)
y3.inf=rbind(y3.inf,m1.inf)
}
x3=rep(n3,4)
plot(x3,y3.inf,type="n",xlab="N",ylab="Pla")

for(i in 1:4){
lines(n3,n3.inf.ans[i,],lty=i,col="red")
}

```

```

}
legend(300,.23,c(".01",".05",".10",".25"),lty=c(1,2,3,4),title="defect
size")
title("large population, binomial (red)")

dev.off()

#minimum sample size for Pla> prob level, infinite population, binomial
counts.p=c(.6,.7,.8,.9) #prob level
counts.3inf=matrix(rep(0,4*4),ncol=4)

for(i in 1:4){
for(j in 1:4){
counts.3inf[i,j]= min(n3[n3.inf.ans[i,]>=counts.p[j]])
}
}
counts.3inf

```

Metric 1b

```

#metric Plb

p.v=c(.01, .05, .10, .25) # problem sizes 1% etc.

pop.v=c(100,500) #population sizes

n1=seq(1,50,1) #sample size 1-50


# infinite population
n1.inf.ans=NULL

y.inf=NULL
for(i in 1:4){
p=p.v[i]
m2.inf=1-(1-p)^n1
n1.inf.ans=rbind(n1.inf.ans,m2.inf)
y.inf=rbind(y.inf,m2.inf)
}


# population 100, treat as finite
n.pop=pop.v[1]
n1a.ans=NULL
for(i in 1:4){
p=p.v[i]
n.p=n.pop*p
m2.small=1-dhyper(n1,n.pop-n.p,n.p,n1)
n1a.ans=rbind(n1a.ans,m2.small)
}

#population 500, treat as finite

```

```

n.pop=pop.v[2]
n1b.ans=NULL
for(i in 1:4){
p=p.v[i]
n.p=n.pop*p
m2.small=1-dhyper(n1,n.pop-n.p,n.p,n1)
n1b.ans=rbind(n1b.ans,m2.small)
}

pdf("metric Plb plots.pdf")
x=rep(n1,4)
plot(x,y.inf,type="n",xlab="N",ylab="Plb")
for(i in 1:4){
lines(n1,n1a.ans[i,],lty=i)
}

for(i in 1:4){
lines(n1,n1b.ans[i,],lty=i,col="green")
}

for(i in 1:4){
lines(n1,n1.inf.ans[i,],lty=i,col="red")
}

legend(25,.60,c(".01",".05",".10",".25"),lty=c(1,2,3,4),title="defect
size")
title("100 (black), 500 (green), large (red) populations")
dev.off()

#minimum sample size for Plb > prob level, population 100
counts.p=c(.6,.7,.8,.9) # probe level
counts.1a=matrix(rep(0,4*4),ncol=4)
for(i in 1:4){
for(j in 1:4){
counts.1a[i,j]= min(n1[n1a.ans[i,]>=counts.p[j]])
}
}
counts.1a

#minimum sample size for Plb > prob level, population 500
counts.1b=matrix(rep(0,4*4),ncol=4)
for(i in 1:4){
for(j in 1:4){
counts.1b[i,j]= min(n1[n1b.ans[i,]>=counts.p[j]])
}
}
counts.1b

#minimum sample size for Plb > prob level, infinite population
counts.1inf=matrix(rep(0,4*4),ncol=4)
for(i in 1:4){
for(j in 1:4){
counts.1inf[i,j]= min(n1[n1.inf.ans[i,]>=counts.p[j]])
}
}

```

```

}
counts.linf

```

Metric 2a

```

#metric P2a
pm.v=c(.75, .90, .95, .99) #targets

p.v=c(.80, .90, .95, .99, .995,.999, .99999) #population proportion above
spec limit

#n=2-50 for normal
#n=2-25 for nonnormal

n.v=seq(1,50,1) # sample size
ans=array(0,c(50,7,4,3))

#normal
mu=0
sigma=1

nsim=10000 #number of simulated data sets

for(i in 2:50){
  n=n.v[i]
  for(j in 1:7){
    p=p.v[j] #true prob
    for(k in 1:4){
      pm=pm.v[k] # metric cutoff
      L= qnorm(1-p)
      nu=n-1
      delta=sqrt(n)*qnorm(pm)
      temp=rep(0,nsim)

      for(isim in 1:nsim){
        y=rnorm(n,0,1)
        ybar=mean(y)
        s=sqrt(var(y))
        arg= (ybar-L)/(s/sqrt(n))

        conf=pt(arg,nu,delta)
        #conf2=max(conf,1-conf)

        #conf2
        #temp[isim]=conf2
        temp[isim]=conf
      }
      ans[i,j,k,]=quantile(temp,c(.10, .50, .90))
      print(c(i,j,k))
    }#k
  }#j
}#i

# target .75 .9 .95 .99 table for n=10, 20, 30, 40, 50

```

```

# true .80, .90, .95, .99, .995,.999, .99999
# .1, .5, .9 quantiles
# table columns n quantile true .80, .90, .95, .99, .995,.999, .99999
for(getind in 1:4){
print(c("target: ",pm.v[getind]))
new.tab=NULL
row.lab=NULL
for(n in seq(10,50,10)){
#print(n)
new.tab=rbind(new.tab, round(t(ans[n,,getind,]),3) )
row.lab=rbind(row.lab,cbind(n,c(.1,.5,.9)))
}
print("n quantile true .80, .90, .95, .99, .995,.999, .99999")
print(cbind(row.lab,new.tab))
}

```

```
pdf("metric P2a true normal.pdf")
```

```

for(j in 1:7){
for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)
y=c(t(ans[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
lines(x,ans[2:50,j,k,1],lty=2)
lines(x,ans[2:50,j,k,2],lty=1)
lines(x,ans[2:50,j,k,3],lty=2)
title(paste(".1, .5, .9 quantiles, normal: true ", p.v[j], ", target
",pm.v[k],sep=""))

}
}

```

```
dev.off()
```

```
#combine previous plots
```

```

pdf("metric P2a true normal combined.pdf")
col.v=c("black","red","green","blue","brown","blueviolet","deeppink")
#{lines(c(-1,0),c(.1*i,.1*i),col=col.v[i])}

```

```

for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)
y=c(t(ans[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
for(j in 1:7){
lines(x,ans[2:50,j,k,1],lty=2,col=col.v[j])

```

```

lines(x,ans[2:50,j,k,2],lty=1,col=col.v[j])
#lines(x,ans[2:50,j,k,3],lty=2,col=col.v[j])
}
#title(paste(".1(d), .5(s), .9(d) quantiles, normal, target
",pm.v[k],sep=""))
title(paste(".1(dashed), .5(solid) quantiles, normal, target
",pm.v[k],sep=""))

mtext("True: .8(blk), .9(r), .95(g), .99(blu), .995(br), .999(v),
.99999(p)",3,0)

}

dev.off()

#target .9
# minimum sample size that give quantile exceeds prob level
q.v=c(.1,.5,.9)
for(ll in 1:3){# ll 1-3 for the .10, .5, .9 quantiles

counts.p=c(.6,.7,.8,.9)#prob level
counts.3=NULL

for(j in 3:7){ #true p
for(k in 2:2){ #p target .9

pTrue=p.v[j]
pTarget=pm.v[k]

temp=rep(0,4)
for(oo in 1:4){
temp[oo]= min(n.v[ans[,j,k,ll]>=counts.p[oo]]) #0.50 quantile
}

counts.3=rbind(counts.3,c(pTrue,pTarget,temp))
}
}
print(c("ll, q.v[ll]: ",ll,q.v[ll]))
print(counts.3)
}#ll

#do all targets
# minimum sample size that give quantile exceeds prob level
q.v=c(.1,.5,.9)

for(k in 1:4){#
pTarget=pm.v[k]

for(ll in 1:3){# ll 1-3 for the .10, .5, .9 quantiles

counts.p=c(.6,.7,.8,.9)#prob level

```

```

counts.3=NULL

if(k==1) start=1
if(k!=1) start=k+1
for(j in start:7){ #true p

pTrue=p.v[j]

temp=rep(0,4)
for(oo in 1:4){
temp[oo]= min(n.v[ans[,j,k,ll]>=counts.p[oo]]) #0.50 quantile
}

counts.3=rbind(counts.3,c(pTrue,pTarget,temp))
}#j

print(c("pTarget, q.v[ll]: ",pTarget,q.v[ll]))
print(counts.3)
}#ll
}#k

```

```

#lognormal - same sigma as normal

```

```

n.v=seq(1,50,1)
ans.ln1=array(0,c(50,7,4,3))

mu=0
sigma=sqrt(.481)
nsim=10000 #sample size

for(i in 2:50){
n=n.v[i]
for(j in 1:7){
p=p.v[j] #true prob
for(k in 1:4){
pm=pm.v[k] # metric cutoff
L.ln=-exp(qnorm(p,0,sqrt(.481)))
nu=n-1
delta=sqrt(n)*qnorm(pm)
temp=rep(0,nsim)

for(isim in 1:nsim){
y= -exp(rnorm(n,0,sqrt(.481)))
ybar=mean(y)
s=sqrt(var(y))
arg= (ybar-L.ln)/(s/sqrt(n))

conf=pt(arg,nu,delta)
#conf2=max(conf,1-conf)
#conf2

```

```

#temp[isim]=conf2
temp[isim]=conf
}
ans.ln1[i,j,k,]=quantile(temp,c(.10, .50, .90))
print(c(i,j,k))
}#k
}#j
}#i

#lognormal - larger sigma as normal about 5 times larger

n.v=seq(1,50,1)
ans.ln2=array(0,c(50,7,4,3))

mu=0
sigma=1
nsim=10000 #sample size

for(i in 2:50){
n=n.v[i]
for(j in 1:7){
p=p.v[j] #true prob
for(k in 1:4){
pm=pm.v[k] # metric cutoff
L.ln=-exp(qnorm(p,0,1))
nu=n-1
delta=sqrt(n)*qnorm(pm)
temp=rep(0,nsim)

for(isim in 1:nsim){
y=-exp(rnorm(n,0,1))
ybar=mean(y)
s=sqrt(var(y))
arg= (ybar-L.ln)/(s/sqrt(n))

conf=pt(arg,nu,delta)
#conf2=max(conf,1-conf)
#conf2
#temp[isim]=conf2
temp[isim]=conf
}
ans.ln2[i,j,k,]=quantile(temp,c(.10, .50, .90))
print(c(i,j,k))
}#k
}#j
}#i

pdf("metric P2a lognormal vs normal.pdf")

for(j in 1:7){
for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)

```



```

y=c(t(ans.ln1[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
mtext("lognormal same variance (black), normal (red)", 3, 0)
lines(x,ans.ln1[2:50,j,k,1],lty=2)
lines(x,ans.ln1[2:50,j,k,2],lty=1)
lines(x,ans.ln1[2:50,j,k,3],lty=2)
title(paste(".1, .5, .9 quantiles, lognormal: true ", p.v[j], ", target
",pm.v[k],sep=""))
lines(x,ans[2:50,j,k,1],lty=2,col="red")
lines(x,ans[2:50,j,k,2],lty=1,col="red")
lines(x,ans[2:50,j,k,3],lty=2,col="red")
}
}

for(j in 1:7){
for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)
y=c(t(ans.ln2[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
mtext("lognormal large variance (black), normal (red)", 3, 0)
lines(x,ans.ln2[2:50,j,k,1],lty=2)
lines(x,ans.ln2[2:50,j,k,2],lty=1)
lines(x,ans.ln2[2:50,j,k,3],lty=2)
title(paste(".1, .5, .9 quantiles, lognormal: true", p.v[j], ", target
",pm.v[k],sep=""))
lines(x,ans[2:50,j,k,1],lty=2,col="red")
lines(x,ans[2:50,j,k,2],lty=1,col="red")
lines(x,ans[2:50,j,k,3],lty=2,col="red")
#legend() #ln2 vs normal
}
}

for(j in 1:7){
for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)
y=c(t(ans.ln1[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
mtext("lognormal: same variance (black), large variance (red)", 3, 0)
lines(x,ans.ln1[2:50,j,k,1],lty=2)
lines(x,ans.ln1[2:50,j,k,2],lty=1)
lines(x,ans.ln1[2:50,j,k,3],lty=2)
title(paste(".1, .5, .9 quantiles, lognormal true ", p.v[j], ", target
",pm.v[k],sep=""))
lines(x,ans.ln2[2:50,j,k,1],lty=2,col="red")
lines(x,ans.ln2[2:50,j,k,2],lty=1,col="red")
lines(x,ans.ln2[2:50,j,k,3],lty=2,col="red")
#legend() #ln1 vs ln2
}
}

```

```

for(j in 1:7){
for(k in 1:4){

x=n.v[-1]
xx=rep(n.v[-1],3)
y=c(t(ans.ln1[2:50,j,k,]))
plot(xx,y,type="n",xlab="N",ylab="P2a",ylim=c(0,1))
mtext("lognormal: same variance (black), large variance (red); normal
(green)", 3, 0)
lines(x,ans.ln1[2:50,j,k,1],lty=2)
lines(x,ans.ln1[2:50,j,k,2],lty=1)
lines(x,ans.ln1[2:50,j,k,3],lty=2)
title(paste(".1, .5, .9 quantiles, lognormal true ", p.v[j], ", target
",pm.v[k], sep=""))
lines(x,ans.ln2[2:50,j,k,1],lty=2,col="red")
lines(x,ans.ln2[2:50,j,k,2],lty=1,col="red")
lines(x,ans.ln2[2:50,j,k,3],lty=2,col="red")
#legend() #ln1 vs ln2
lines(x,ans[2:50,j,k,1],lty=2,col="green")
lines(x,ans[2:50,j,k,2],lty=1,col="green")
lines(x,ans[2:50,j,k,3],lty=2,col="green")

}
}

dev.off()

pdf("compare metric P2a lognormal vs normal pdfs.pdf")

y1=-exp(rnorm(100000,0,sqrt(.481)))
y2=-exp(rnorm(100000,0,1))
plot(density(y1),main="LN same var (black), normal (red), LN larger var
(green)",xlim=c(-20,3),xlab="")
x1=seq(-3,3,.1)
lines(x1-1.27,dnorm(x1),lty=1, col="red")
lines(density(y2),lty=1,col="green")
dev.off()

# comment out two following commands
#save.image("save metric 3 calculations in R revised 041811") # save
results to replot later
#load("save metric 3 calculations in R revised 041811") # load results to
replot

zz=exp(rnorm(100000,0,sqrt(.481)))
var(zz)
mean(zz)

> var(zz)
[1] 0.995539
> mean(zz)
[1] 1.270782
plot(0,0)
col.v=c("black","red","green","blue","brown","blueviolet","deeppink")

```

```

for(i in 1:7){
{lines(c(-1,0),c(.1*i,.1*i),col=col.v[i])}

#if(i!=7){lines(c(-1,0),c(.1*i,.1*i),col=col.v[i])}
#if(i==7){i=9;lines(c(-1,0),c(.1*i,.1*i),col=i)}
}

```

Metric 2b/2c

```

#metric P3b, P3c

# first predict at end of period where end of period probabilities below
spec limit are .01, .05, .10, .25
#      this is to compare with metric P2a; this is metric P2b

# then predict 15 years later where 15 year later probabilities below spec
limit are .01, .05, .10, .25, slopes
# are necessarily less in second scenario
# this is metric P2c

# for now let's do even spacing and .90 case

#metric P2b
n.v=seq(10,100,10) # sample sizes in increments of 10
pBegin.v=c(.80, .90, .95, .99, .995, .999, .9999, .999999) #proportions
above spec L at time 0
pEnd.v=c(.75, .90, .95, .99, .995, .999, .9999) #proportions above spec L
at now or 15 years in future

ans.5a=array(0,c(8,7, 2,10, 3,3))
m.v=c(1,2,3)
per.v=c(10,30)

#m.v values - only use equal spaced m.v[1]
#1 equal spaced
#2 50:50 at begin (year 1) and end
#3 80:20 at begin (year 1) and end

nsim=1000 # number of simulated data sets
pTarget=.9 # target 90% above spec limit

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

for(k in 1:2){
for(ll in 1:10){
for(mm in 1:1){

pBegin=pBegin.v[i]
pEnd=pEnd.v[j]
per=per.v[k]
n=n.v[ll]
m=m.v[mm]

```

```

L=0 # spec limit 0, without loss of generality

if(pEnd<pBegin){
print(c("i,j,k,ll,mm: ",i,j,k,ll,mm))
beta0OverSigma=qnorm(pBegin)
kEnd=qnorm(pEnd)

beta1OverSigma=((kEnd-beta0OverSigma)/(per)) # at end of period

if(m==1){#1
if(per==10){nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(per==30){#2
rem=n%%per;
if(rem==0){nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(rem!=0){#3
nRep=floor(n/per);
x=rep(seq(1,per,1),nRep);
if(rem==10) x=c(x,seq(3,30,3))
if(rem==20) x=c(x,seq(2,29,3), seq(3,30,3))
}#3
}#2
}#1
if(m==2) x=c(rep(1,n/2),rep(per,n/2))
if(m==3) x=c(rep(1,n*.8),rep(per,n*.2))
x=sort(x) #put in age order

X=cbind(1,x)
xt=c(1,per)
H=xt%%solve(t(X)%%X)%%cbind(xt)

nu=n-2

delta=qnorm(pTarget)/sqrt(H)
mu=beta0OverSigma+beta1OverSigma*x
mu.per=beta0OverSigma+beta1OverSigma*per

temp=rep(0,nsim)
for(isim in 1:nsim){
y=mu+rnorm(n)
ls.fit=summary(lm(y~x))
s=ls.fit$sigma
ypred=xt%%cbind(ls.fit$coef[,1])
arg= (ypred-L)/(s*sqrt(H))

conf=pt(arg,nu,delta)
temp[isim]=conf
}

ans.5a[i,j,k,ll,mm,]=quantile(temp,c(.1,.5,.9))
}#if

}#mm
}#ll

```

```

}#k
}#j
}#i

# minimum sample size for P2b > prob level
counts.p=c(.6,.7,.8,.9) #prob level
counts.5a=NULL

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

for(k in 1:2){
pBegin=pBegin.v[i]
pEnd=pEnd.v[j]
per=per.v[k]

temp=rep(0,4)
for(oo in 1:4){
temp[oo]= min(n.v[ans.5a[i,j,k,,1,3]>=counts.p[oo]]) #0.50 quantile
}

counts.5a=rbind(counts.5a,c(pBegin,pEnd,.9,per,temp))
}
}
}
counts.5a

pdf("metric P2b normal.pdf")

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

x=n.v
xx=rep(n.v,6)
y=c(ans.5a[i,j,1,,1,1],ans.5a[i,j,1,,1,2],ans.5a[i,j,1,,1,3],ans.5a[i,j,2,
,1,1],ans.5a[i,j,2,,1,2],ans.5a[i,j,2,,1,3])
plot(xx,y,type="n",xlab="N",ylab="P2b",ylim=c(0,1))
lines(x,ans.5a[i,j,1,,1,1],lty=2)
lines(x,ans.5a[i,j,1,,1,2],lty=1)
lines(x,ans.5a[i,j,1,,1,3],lty=2)
lines(x,ans.5a[i,j,2,,1,1],lty=2,col="red")
lines(x,ans.5a[i,j,2,,1,2],lty=1,col="red")
lines(x,ans.5a[i,j,2,,1,3],lty=2,col="red")
title(paste(".1, .5, .9 quantiles, equal spacing, normal"))
mtext(paste(" start ", pBegin.v[i], ", end (per) ",pEnd.v[j], ", target
0.9",sep=""),3,.75)
mtext("period: 10 (black), 30 (red); prediction at end of period",3,0)

}
}

dev.off()

```

```

#metric P2c
n.v=seq(10,100,10)
pBegin.v=c(.80, .90, .95, .99, .995, .999, .9999, .999999)
pEnd.v=c(.75, .90, .95, .99, .995, .999, .9999)
ans.5b=array(0,c(8,7, 2,10, 3,3))
m.v=c(1,2,3)
per.v=c(10,30)

#m.v values - only use equal spaced m.v[1]
#1 equal spaced
#2 50:50 at begin (year 1) and end
#3 80:20 at begin (year 1) and end

nsim=1000 # number of simulated data sets
pTarget=.9 # target 90% above spec limit

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd
for(k in 1:2){
for(ll in 1:10){
for(mm in 1:1){

pBegin=pBegin.v[i]
pEnd=pEnd.v[j]
per=per.v[k]
n=n.v[ll]
m=m.v[mm]

L=0 # spec limit 0, without loss of generality

if(pEnd<pBegin){
print(c("i,j,k,ll,mm: ",i,j,k,ll,mm))
beta0OverSigma=qnorm(pBegin)
kEnd=qnorm(pEnd)

beta1OverSigma=((kEnd-beta0OverSigma)/(per+15)) # at 15 years past end of
period

if(m==1){#1
if(per==10) {nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(per==30){#2
rem=n%%per;
if(rem==0) {nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(rem!=0){#3
nRep=floor(n/per);
x=rep(seq(1,per,1),nRep);
if (rem==10) x=c(x,seq(3,30,3))
if(rem==20) x=c(x,seq(2, 29, 3), seq(3,30,3))
}#3
}#2
}#1

```

```

if(m==2) x=c(rep(1,n/2),rep(per,n/2))
if(m==3) x=c(rep(1,n*.8),rep(per,n*.2))
x=sort(x) # sort by age

X=cbind(1,x)
xt=c(1,(per+15)) # predict 15 years past end of period
H=xt%%solve(t(X)%%X)%%cbind(xt)

nu=n-2

delta=qnorm(pTarget)/sqrt(H)
mu=beta0OverSigma+beta1OverSigma*x
mu.per=beta0OverSigma+beta1OverSigma*per

temp=rep(0,nsim)
for(isim in 1:nsim){
y=mu+rnorm(n)
ls.fit=summary(lm(y~x))
s=ls.fit$sigma
ypred=xt%%cbind(ls.fit$coef[,1])
arg= (ypred-L)/(s*sqrt(H))

conf=pt(arg,nu,delta)
temp[isim]=conf
}

ans.5b[i,j,k,ll,mm,]=quantile( temp,c(.1,.5,.9))
}#if

}#mm
}#ll
}#k
}#j
}#i

pdf("metric P2c normal.pdf")
for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

x=n.v
xx=rep(n.v,6)
y=c(ans.5b[i,j,1,,1,1],ans.5b[i,j,1,,1,2],ans.5b[i,j,1,,1,3],ans.5b[i,j,2,
,1,1],ans.5b[i,j,2,,1,2],ans.5b[i,j,2,,1,3])
plot(xx,y,type="n",xlab="N",ylab="P2c",ylim=c(0,1))
lines(x,ans.5b[i,j,1,,1,1],lty=2)
lines(x,ans.5b[i,j,1,,1,2],lty=1)
lines(x,ans.5b[i,j,1,,1,3],lty=2)
lines(x,ans.5b[i,j,2,,1,1],lty=2,col="red")
lines(x,ans.5b[i,j,2,,1,2],lty=1,col="red")
lines(x,ans.5b[i,j,2,,1,3],lty=2,col="red")
title(paste(".1, .5, .9 quantiles, equal spacing, normal"))
mtext(paste("start ", pBegin.v[i], ", end (per + 15) ",pEnd.v[j],",target
0.9 ",sep=""),3, .75)

```

```

mtext("period: 10 (black), 30 (red), prediction at end of period +
15",3,0)

}
}
dev.off()

#minimum sample size P2c > prob level
counts.p=c(.6,.7,.8,.9) #prob level
counts.5b=NULL

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

for(k in 1:2){
pBegin=pBegin.v[i]
pEnd=pEnd.v[j]
per=per.v[k]

temp=rep(0,4)
for(oo in 1:4){
temp[oo]= min(n.v[ans.5b[i,j,k,,1,3]>=counts.p[oo]]) #0.50 quantile
}

counts.5b=rbind(counts.5b,c(pBegin,pEnd,.9,per,temp))
}
}
}
counts.5b

# two following commands are commented out
#save.image("save metric 5 calculations in R 041811") #save calculations
if you need them to replot
#load("save metric 5 calculations in R 041811") #load calculations if you
need them to replot

```

Metric 3a/3b

```

#metric p3a

n=seq(1,50,1) #sample size 1-50
m.v=seq(1,3,1) #mean shifts of 1-3 sd

cv=qchisq(.95,n) #0.05 test
ans=NULL
for(i in 1:3){
m=m.v[i]
lambda=n*m^2

ans=rbind(ans,1-pchisq(cv,n,lambda))
}

```



```

pdf("metric P3a P3b plots.pdf")
y=c(t(ans))
x=rep(n,3)
plot(x,y,type="n",xlab="N",ylab="P3a")
for(i in 1:3){
  lines(n,ans[i,],lty=i)
}

legend(15,.5,c("1 sd","2 sd","3 sd"),lty=c(1,2,3,4,5),title="mean shift")

# get min counts for which P3a> prob level
counts.p=c(.6,.7,.8,.9) #prob levels of .6, .7, .8., .9
counts=matrix(rep(0,3*4),ncol=4)

for(i in 1:3){
  for(j in 1:4){
    counts[i,j]= min(n[ans[i,]>=counts.p[j]])
  }
}
counts

#metric P3b

n=seq(1,50,1) #sample size 1-50
m.v=c(1.5,2,3) #sd multipliers 1.5, 2, 3

cv=qchisq(.95,n) #0.05 test
ans=NULL
for(i in 1:3){
  m=m.v[i]

  ans=rbind(ans,1-pchisq((cv/m^2),n))
}

y=c(t(ans))
x=rep(n,3)
plot(x,y,type="n",xlab="N",ylab="P3b")
for(i in 1:3){
  lines(n,ans[i,],lty=i)
}

legend(15,.5,c("1.5","2","3"),lty=c(1,2,3),title="sd multiplier")

# get min counts for which P3b> prob level
counts.p=c(.6,.7,.8,.9) #prob levels of .6, .7, .8., .9
counts=matrix(rep(0,3*4),ncol=4)

for(i in 1:3){
  for(j in 1:4){
    counts[i,j]= min(n[ans[i,]>=counts.p[j]])
  }
}
counts
dev.off()

```

Metric 4

07.26.11

```
#compare line asymptote step half-line for metric 4 robustness study
per=30
#per=10
pBegin=.999
pEnd=.9
x=seq(0,per,1)

beta0OverSigma=qnorm(pBegin)
kEnd=qnorm(pEnd)

beta1OverSigma=((kEnd-beta0OverSigma)/(per+15)) #power formula based on
positive slope

muline=beta0OverSigma+beta1OverSigma*x
plot(x,muline,type="l",xlab="year",ylab="mu")

#asymptote (previous but close to line)
#r=log((beta0OverSigma+beta1OverSigma*per)/beta0OverSigma)/per #k is
period
#muasymptote=beta0OverSigma*exp(r*x)
#lines(x,muasymptote,col="red")

#use alternate asymptote
qEnd=beta0OverSigma+beta1OverSigma*per
qBegin=qnorm(pBegin)
if(per==10) muasymptote=((qBegin-qEnd)/(3*x+1))+qEnd
if(per==30) muasymptote=((qBegin-qEnd)/((x)+1))+qEnd
lines(x,muasymptote,col="red")

lines(c(0,per/2),rep(beta0OverSigma,2),col="green")
lines(c(per/2,per),rep(beta0OverSigma+beta1OverSigma*per,2),col="green")

lines(c(0,per/2),rep(beta0OverSigma+.005,2),col="blue")

beta1OverSigmaHalfline=((beta0OverSigma+beta1OverSigma*per)-
beta0OverSigma)/((per/2))

xx=seq(per/2,per,.5)
mu=beta0OverSigma+beta1OverSigmaHalfline*(xx - (per/2))
lines(xx,mu,col="blue")

title(paste("mean start ",pBegin," end (per+15) ",pEnd," period:
",per,sep=""))
mtext("line (black), asymptote (red), step (green), half-line
(blue)",3,.75)
```

```

#metric 4

#beta0 - intercept
#beta1 - slope
#sigma=1

# initial probabilities above spec are .80, .90, .95, .99, .995, .999,
.9999, .999999
#data time period 10 or 30 years
#data time period + 15 years probabilities above spec are .75, .90, .95,
.99, .995, .999, .9999 - these have to be less than initial probabilities
to
# have a non-zero slope (slope will be negative but make it positive for
the power calculation)

n.v=seq(10,100,10) # sample size 10 to 100 by 10
pBegin.v=c(.80, .90, .95, .99, .995, .999, .9999, .999999)
pEnd.v=c(.75, .90, .95, .99, .995, .999, .9999)
ans=array(0,c(8,7, 2,10, 3))
m.v=c(1,2,3)
per.v=c(10,30)
#1 equal spaced
#2 50:50 at begin (year 1) and end
#3 80:20 at begin (year 1) and end

for(i in 1:8){
  for(j in 1:7){
    for(k in 1:2){
      for(ll in 1:10){
        for(mm in 1:3){

          pBegin=pBegin.v[i]
          pEnd=pEnd.v[j]
          per=per.v[k]
          n=n.v[ll]
          m=m.v[mm]

          if(pEnd<pBegin){
            beta0OverSigma=qnorm(pBegin)
            kEnd=qnorm(pEnd)

            beta1OverSigma=abs((kEnd-beta0OverSigma)/(per+15)) #power formula based on
            positive slope

            if(m==1){#1
              if(per==10) {nRep=n/per; x=rep(seq(1,per,1),nRep)}
              if(per==30){#2
                rem=n%%per;
                if(rem==0) {nRep=n/per; x=rep(seq(1,per,1),nRep)}
                if(rem!=0){#3
                  nRep=floor(n/per);
                  x=rep(seq(1,per,1),nRep);
                  if (rem==10) x=c(x,seq(3,30,3))
                  if(rem==20) x=c(x,seq(2, 29, 3), seq(3,30,3))
                }
              }
            }
          }
        }
      }
    }
  }
}

```

```

}#3
}#2
}#1
if(m==2) x=c(rep(1,n/2),rep(per,n/2))
if(m==3) x=c(rep(1,n*.8),rep(per,n*.2))

nu=n-2
cv=qt(.95,nu)
delta=beta1OverSigma*sqrt(var(x)*(n-1))

ans[i,j,k,ll,mm]=1-pt(cv,nu,delta)

}#if

}#mm
}#ll
}#k
}#j
}#i

pdf("metric P4 normal plots.pdf")

for(i in 1:8){
for(j in 1:7){
print(c(i,j))

pBegin=pBegin.v[i]
pEnd=pEnd.v[j]

if(pEnd<pBegin){

plot(0,0,type="n",xlim=c(0,100),ylim=c(0,1),xlab="N",ylab="P4")
title(paste("start ",pBegin," end (per+15) ",pEnd," period: 10 (black),
30 (red)",sep=""))
mtext("even (solid), 80:20 (dotted), 50:50 (dashed)",3,0)
lines(n.v,ans[i,j,1, ,1],lty=1)# per 10 data 1
lines(n.v,ans[i,j,1, ,2],lty=2)# per 10 data 2
lines(n.v,ans[i,j,1, ,3],lty=3)# per 10 data 3
lines(n.v,ans[i,j,2, ,1],lty=1,col="red")# per 30 data 1
lines(n.v,ans[i,j,2, ,2],lty=2,col="red")# per 30 data 2
lines(n.v,ans[i,j,2, ,3],lty=3,col="red")# per 30 data 3
}

}#j
}#i
dev.off()

#minimum sample size (increments of 10) for P4 > prob level
counts.p=c(.6,.7,.8,.9) # prob levels
counts.4=NULL

for(j in 3:7){ #pEnd > 0.90
for(i in (j+1):8){ #pBegin > pEnd

```

```

for(k in 1:2){
  for(m in 1:3){
    pBegin=pBegin.v[i]
    pEnd=pEnd.v[j]
    per=per.v[k]

    temp=rep(0,4)
    for(oo in 1:4){
      temp[oo]= min(n.v[ans[i,j,k,,m]>=counts.p[oo]]) #0.50 quantile
    }

    counts.4=rbind(counts.4,c(pBegin,pEnd,per,m,temp))
  }
}
counts.4

#####
#####
# do some for 4 year period
# multiples of 4 to 100

n.v=seq(4,100,4)
pBegin.v=c(.80, .90, .95, .99,.995, .999, .9999, .999999)
pEnd.v=c(.75, .90, .95, .99,.995, .999, .9999)
ans.4=array(0,c(8,7, 1,25, 3))
m.v=c(1,2,3)
per.v=c(4)
#1 equal spaced
#2 50:50 at begin (year 1) and end
#3 80:20 at begin (year 1) and end

for(i in 1:8){
  for(j in 1:7){
    for(k in 1:1){
      for(ll in 1:25){
        for(mm in 1:3){

          pBegin=pBegin.v[i]
          pEnd=pEnd.v[j]
          per=per.v[k]
          n=n.v[ll]
          m=m.v[mm]

          if(pEnd<pBegin){
            beta0OverSigma=qnorm(pBegin)
            kEnd=qnorm(pEnd)

            beta1OverSigma=abs((kEnd-beta0OverSigma)/(per+15)) #power formula based on
            positive slope

            if(m==1) x=rep(c(1,2,3,4),n/4)
            if(m==2) x=c(rep(1,n/2),rep(per,n/2))

```

```

if(m==3) x=c(rep(1,n*.75),rep(per,n*.25))

nu=n-2
cv=qt(.95,nu)
delta=beta1OverSigma*sqrt(var(x)*(n-1))

ans.4[i,j,k,ll,mm]=1-pt(cv,nu,delta)

}#if

}#mm
}#ll
}#k
}#j
}#i

pdf("metric P4 normal 4 years plots.pdf")

for(i in 1:8){
for(j in 1:7){
print(c(i,j))

pBegin=pBegin.v[i]
pEnd=pEnd.v[j]

if(pEnd<pBegin){

plot(0,0,type="n",xlim=c(0,100),ylim=c(0,1),xlab="N",ylab="P4")
title(paste("start ",pBegin," end (per+15) ",pEnd," period: 4",sep=""))
mtext("even (solid), 75:25 (dotted), 50:50 (dashed)",3,0)
lines(n.v,ans.4[i,j,1, ,1],lty=1)# per 10 data 1
lines(n.v,ans.4[i,j,1, ,2],lty=2)# per 10 data 2
lines(n.v,ans.4[i,j,1, ,3],lty=3)# per 10 data 3

}

}#j
}#i
dev.off()

#####
####

#metric 4 robustness to
# a) asymptote - to same level by end of the period as above -
beta1OverSigma*period
# b) step function - half the period - to same level by end of the period
as above

# do just for 0.90 results

n.v=seq(10,100,10)

```

```

pBegin.v=c(.80, .90, .95, .99, .995,.999, .9999, .999999)
pEnd.v=c( .75, .90, .95, .99,.995, .999, .9999)
ans.asymp=array(0,c(8,7, 2,10, 3))
ans.step=array(0,c(8,7, 2,10, 3))
ans.halfline=array(0,c(8,7, 2,10, 3))

m.v=c(1,2,3)
per.v=c(10,30)
#1 equal spaced
#2 50:50 at begin (year 1) and end
#3 80:20 at begin (year 1) and end

nsim=10000

for(i in 1:8){
for(j in 2:2){ #only .90 at per+15
for(k in 1:2){
for(ll in 1:10){
for(mm in 1:1){ # only equal spacing

pBegin=pBegin.v[i]
pEnd=pEnd.v[j]
per=per.v[k]
n=n.v[ll]
m=m.v[mm]

if(pEnd<pBegin){#if 0
print(c("i,j,k,ll,mm: ",i,j,k,ll,mm))

beta0OverSigma=qnorm(pBegin)
kEnd=qnorm(pEnd)

beta1OverSigma=((kEnd-beta0OverSigma)/(per+15))

if(m==1){#1
if(per==10){nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(per==30){#2
rem=n%%per;
if(rem==0){nRep=n/per; x=rep(seq(1,per,1),nRep)}
if(rem!=0){#3
nRep=floor(n/per);
x=rep(seq(1,per,1),nRep);
if (rem==10) x=c(x,seq(3,30,3))
if(rem==20) x=c(x,seq(2, 29, 3), seq(3,30,3))
}#3
}#2
}#1
if(m==2) x=c(rep(1,n/2),rep(per,n/2))
if(m==3) x=c(rep(1,n*.8),rep(per,n*.2))
x=sort(x) #need to be sorted for step case as well as other cases

nu=n-2
cv=qt(.95,nu)

```

```

# simulate asymp
cnt=rep(0,nsim)

# 1/x form
qEnd=beta0OverSigma+beta1OverSigma*per
qBegin=qnorm(pBegin)
if(per==10) mu=((qBegin-qEnd)/(3*x+1))+qEnd
if(per==30) mu=((qBegin-qEnd)/((x)+1))+qEnd

for(isim in 1:nsim){
y=mu+rnorm(n)
lm.fit=summary(lm(y~x))
tvalue=lm.fit$coef[2,3]
if(tvalue<=-cv) cnt[isim]=1
}
ans.asymp[i,j,k,ll,mm]=sum(cnt)/nsim
#ans.asymp[i,j,k,ll,mm]

#simulate step
cnt=rep(0,nsim)

mu=x
mu[x<=(per/2)]=beta0OverSigma
mu[x>(per/2)]=beta0OverSigma+beta1OverSigma*per
for(isim in 1:nsim){
y=mu+rnorm(n)
lm.fit=summary(lm(y~x))
tvalue=lm.fit$coef[2,3]
if(tvalue<=-cv) cnt[isim]=1
}
ans.step[i,j,k,ll,mm]=sum(cnt)/nsim
#ans.step[i,j,k,ll,mm]

#simulate halfline
cnt=rep(0,nsim)

beta1OverSigmaHalfline=(qEnd-beta0OverSigma)/(per/2) #line start halfway
through period
mu=beta0OverSigma+beta1OverSigmaHalfline*(x - (per/2))
mu[x<=(per/2)]=beta0OverSigma

for(isim in 1:nsim){
y=mu+rnorm(n)
lm.fit=summary(lm(y~x))
tvalue=lm.fit$coef[2,3]
if(tvalue<=-cv) cnt[isim]=1
}
ans.halfline[i,j,k,ll,mm]=sum(cnt)/nsim
#[i,j,k,ll,mm]

}#if 0

}#mm

```



```

}#l1
}#k
}#j
}#i

#make plots
#pdf("metric 4 robustness plots 071811.pdf")
pdf("metric 4 robustness plots 072611.pdf")

for(i in 3:8){
  for(j in 2:2){
    print(c(i,j))

    pBegin=pBegin.v[i]
    pEnd=pEnd.v[j]

    if(pEnd<pBegin){

      plot(0,0,type="n",xlim=c(0,100),ylim=c(0,1),xlab="N",ylab="P4")
      title(paste("equal spacing start ",pBegin," end (per+15) ",pEnd, ",
        period: 10",sep=""))
      mtext("line (black), asymptote (red), step (green), half-line
        (blue)",3,.75)
      lines(n.v,ans[i,j,1, ,1],lty=1)# per 10 data 1
      lines(n.v,ans.asymp[i,j,1, ,1],lty=1,col="red")# per 10 data 1
      lines(n.v,ans.step[i,j,1, ,1],lty=1,col="green")# per 10 data 1
      lines(n.v,ans.halfline[i,j,1, ,1],lty=1,col="blue")# per 10 data 1

      plot(0,0,type="n",xlim=c(0,100),ylim=c(0,1),xlab="N",ylab="P4")
      title(paste("equal spacing start ",pBegin," end (per+15) ",pEnd, ",
        period: 30",sep=""))
      mtext("line (black), asymptote (red), step (green), half-line
        (blue)",3,.75)
      lines(n.v,ans[i,j,2, ,1],lty=1)# per 10 data 1
      lines(n.v,ans.asymp[i,j,2, ,1],lty=1,col="red")# per 10 data 1
      #lines(n.v,ans.asymp[i,j,2, ,2],lty=2,col="red")# per 10 data 2
      #lines(n.v,ans.asymp[i,j,2, ,3],lty=3,col="red")# per 10 data 3
      lines(n.v,ans.step[i,j,2, ,1],lty=1,col="green")# per 10 data 1
      lines(n.v,ans.halfline[i,j,2, ,1],lty=1,col="blue")# per 10 data 1
    }

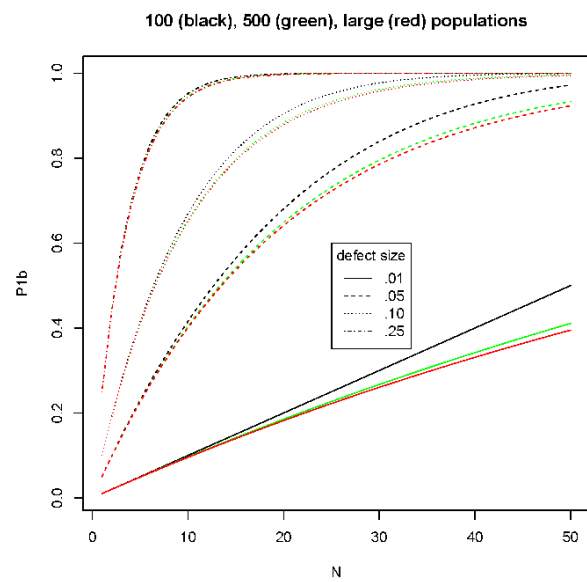
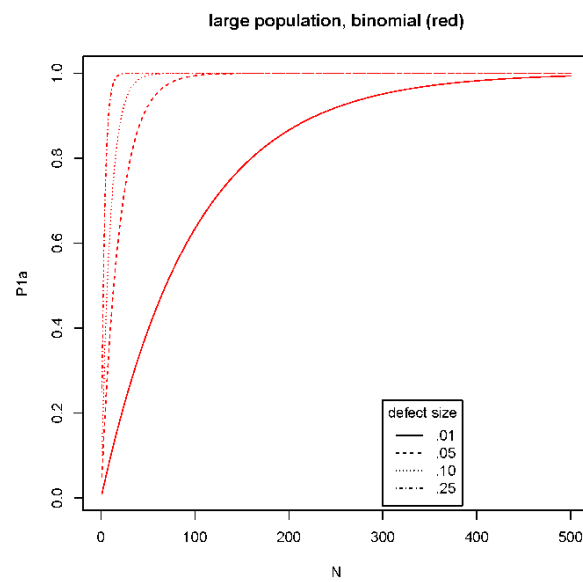
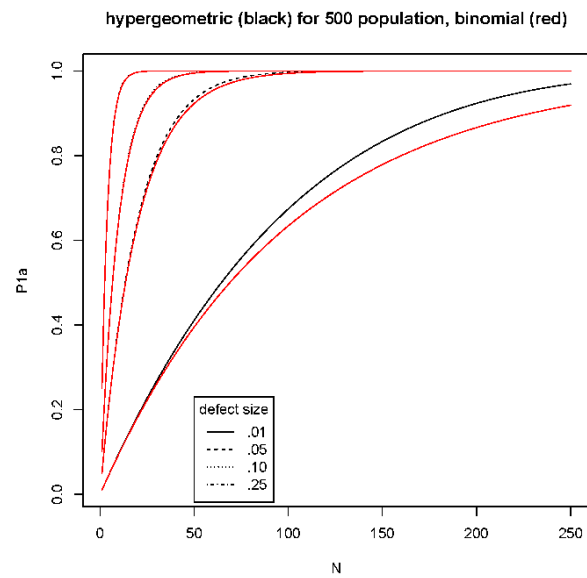
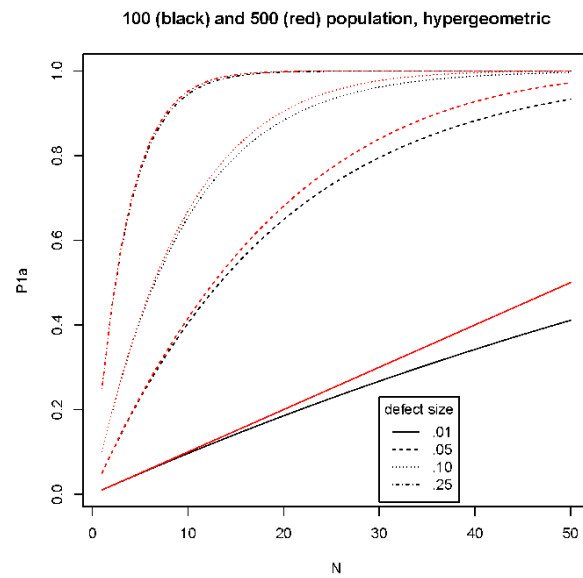
  }#j
}#i
dev.off()

#save.image("save metric 4 calculations in R 072611")
#load("save metric 4 calculations in R 072611")

#save.image("save metric 4 calculations in R 071811")
#load("save metric 4 calculations in R 071811")

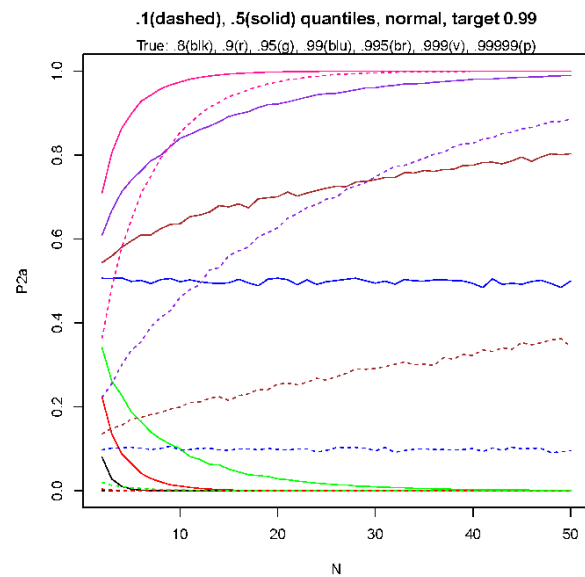
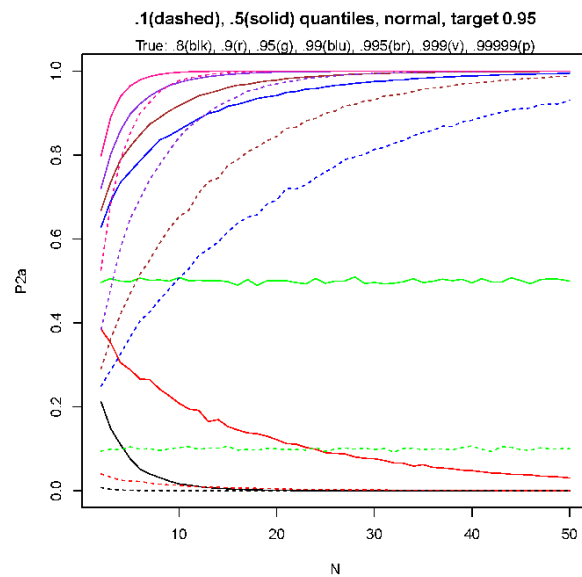
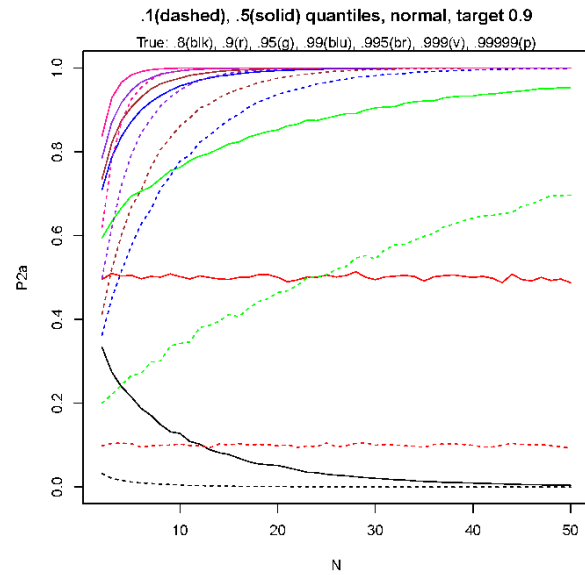
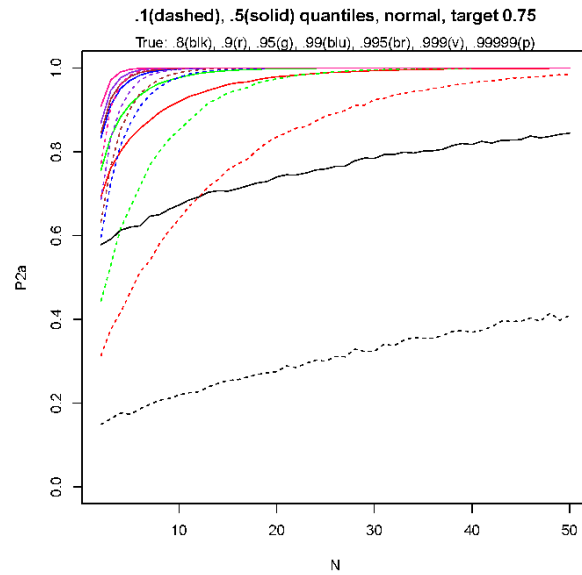
```

APPENDIX B: METRIC 1A AND 1B SENSITIVITY STUDY GRAPHS

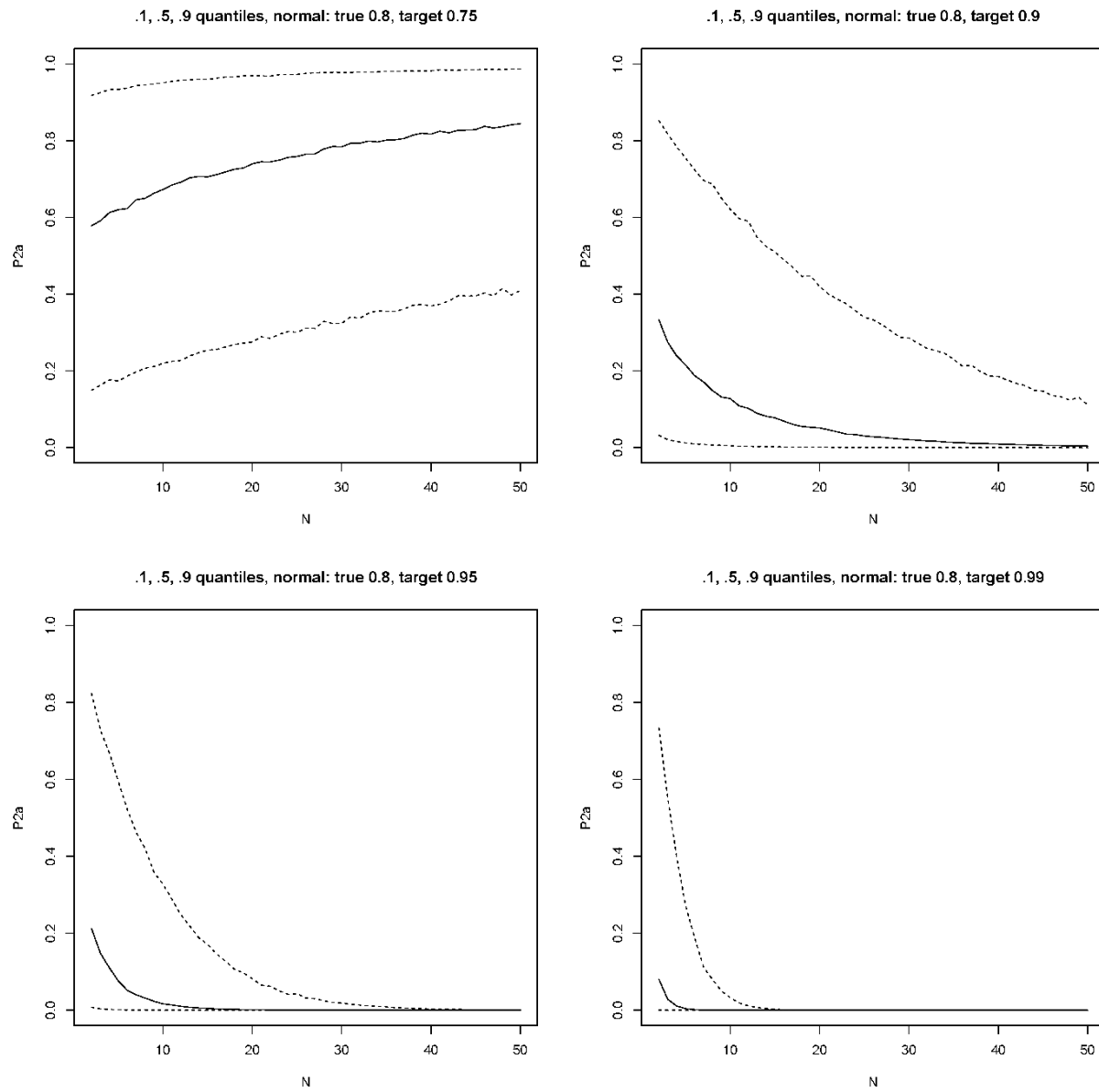


APPENDIX C: METRIC 2A SENSITIVITY STUDY GRAPHS, NORMAL DISTRIBUTION

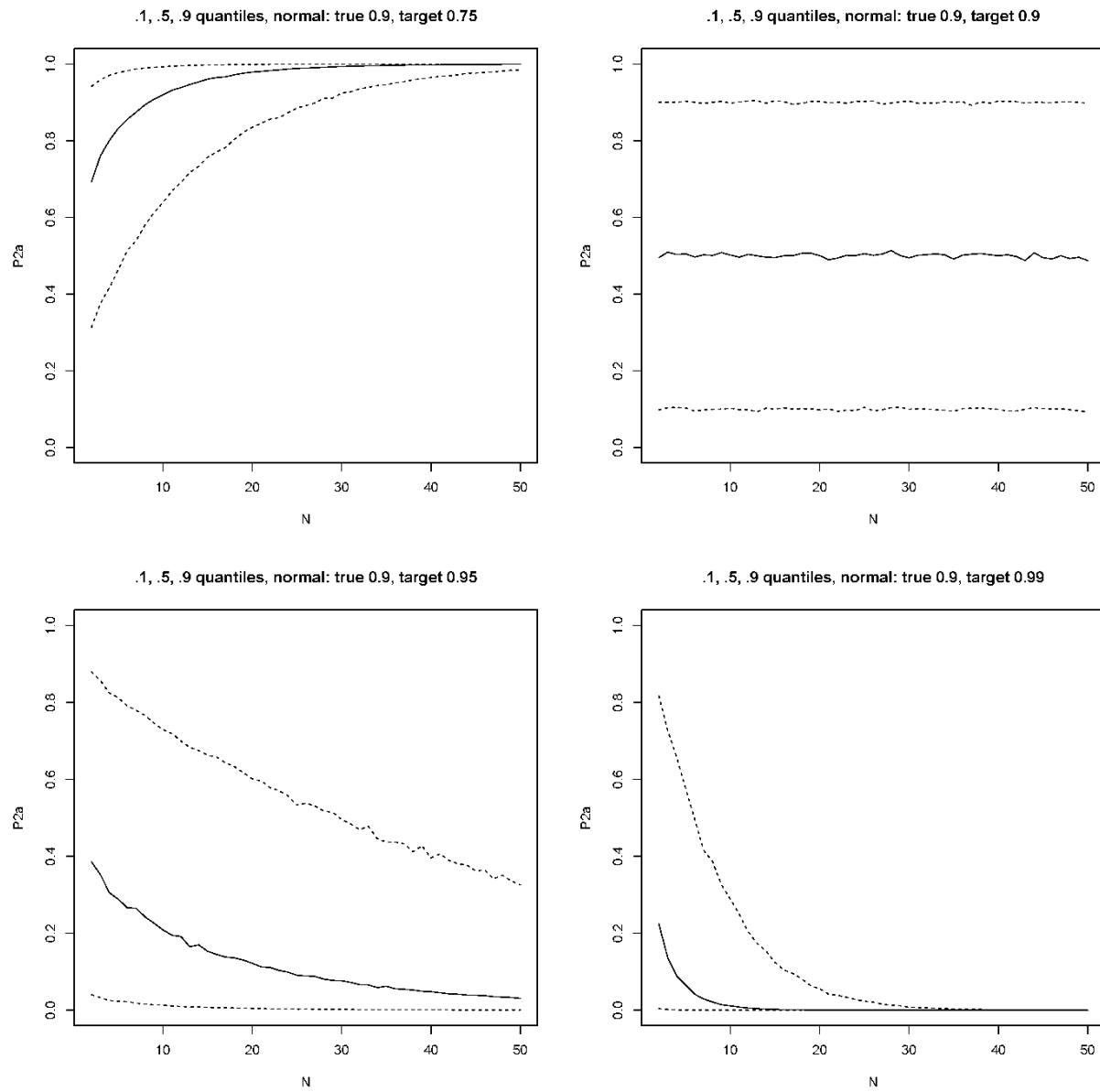
Consolidated Results, Fixed Target Values



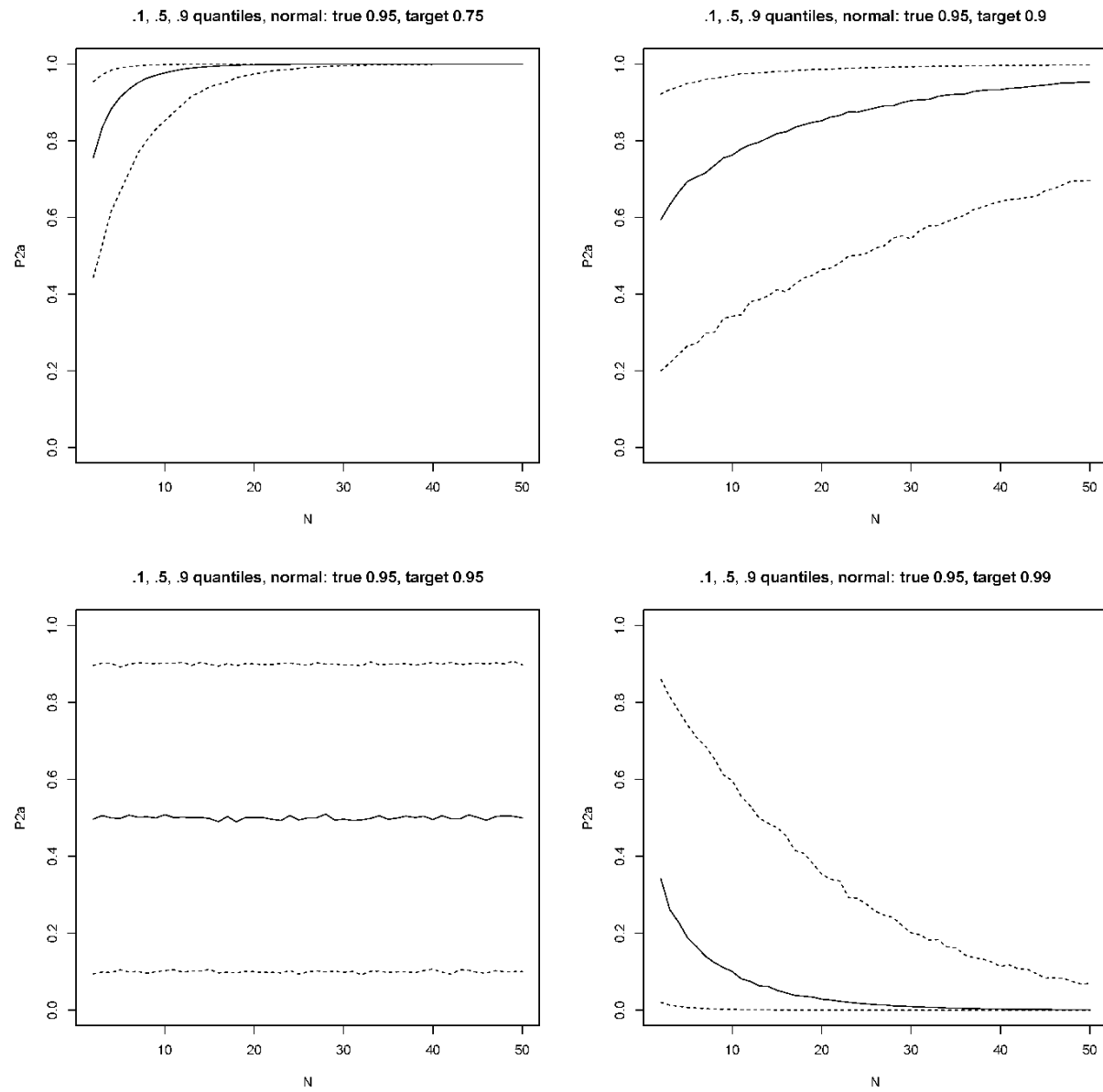
P2a Results, True=0.8, Varying Target



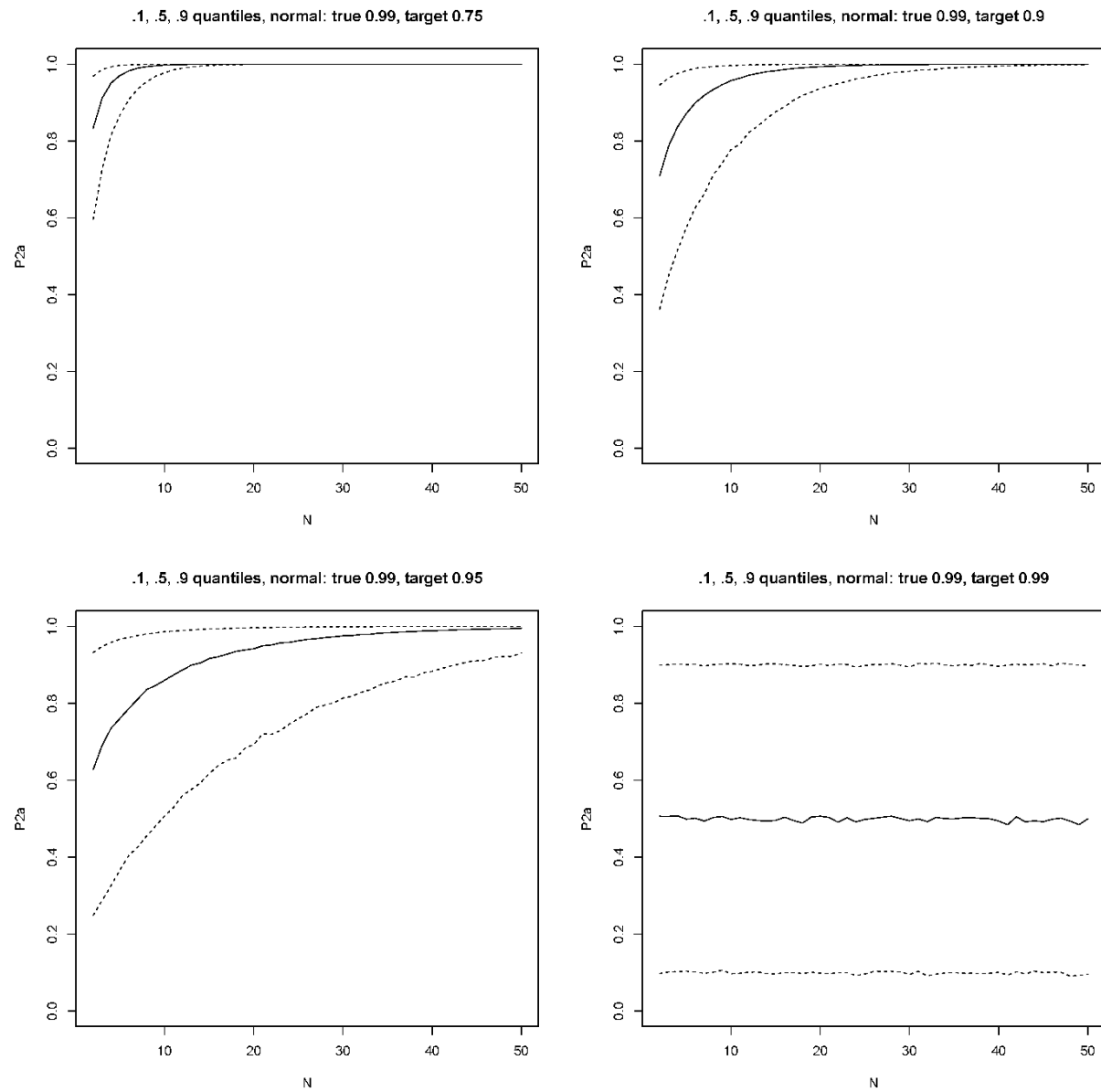
P2a Results, True=0.9, Varying Target



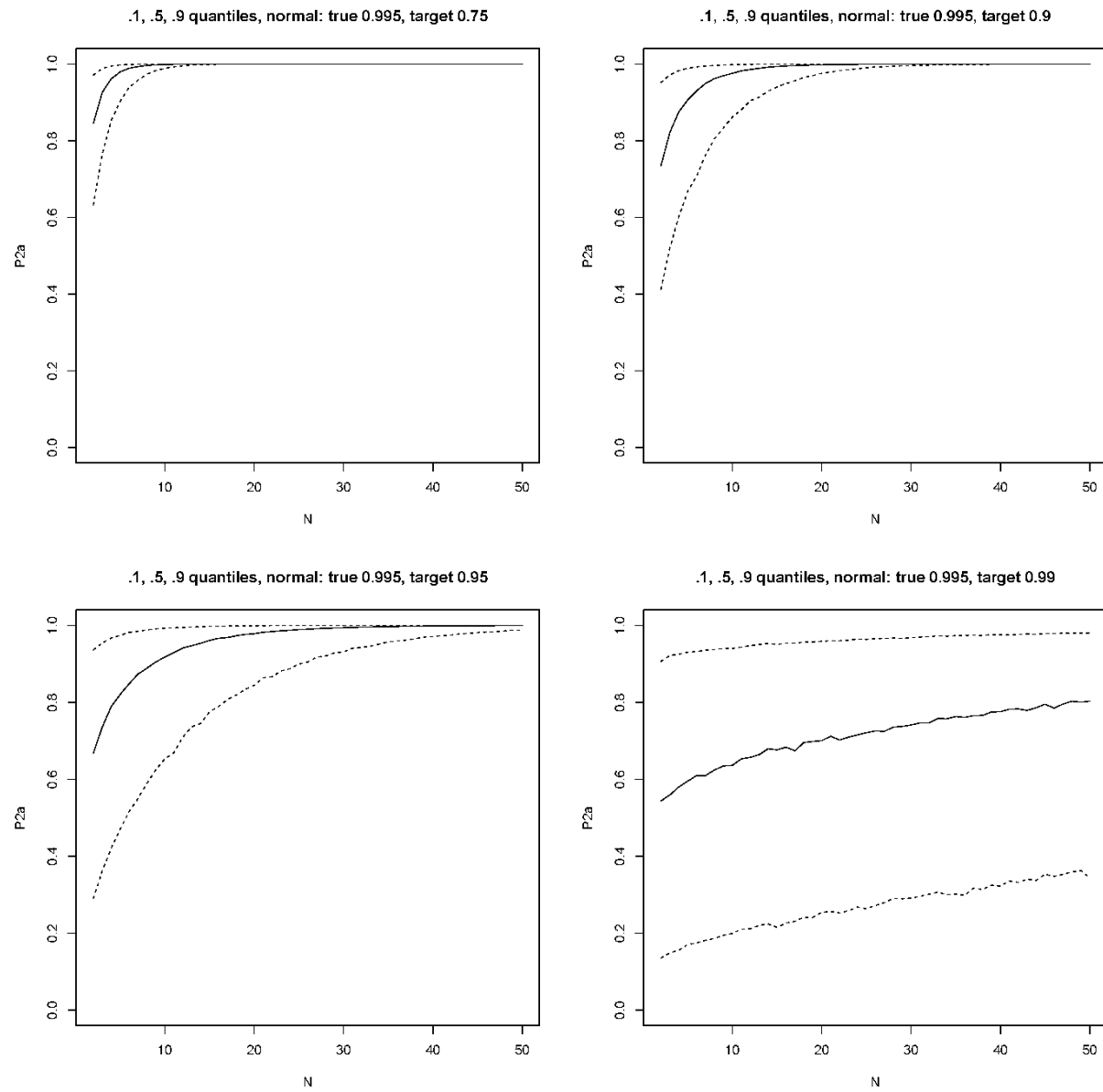
P2a Results, True=0.95, Varying Target



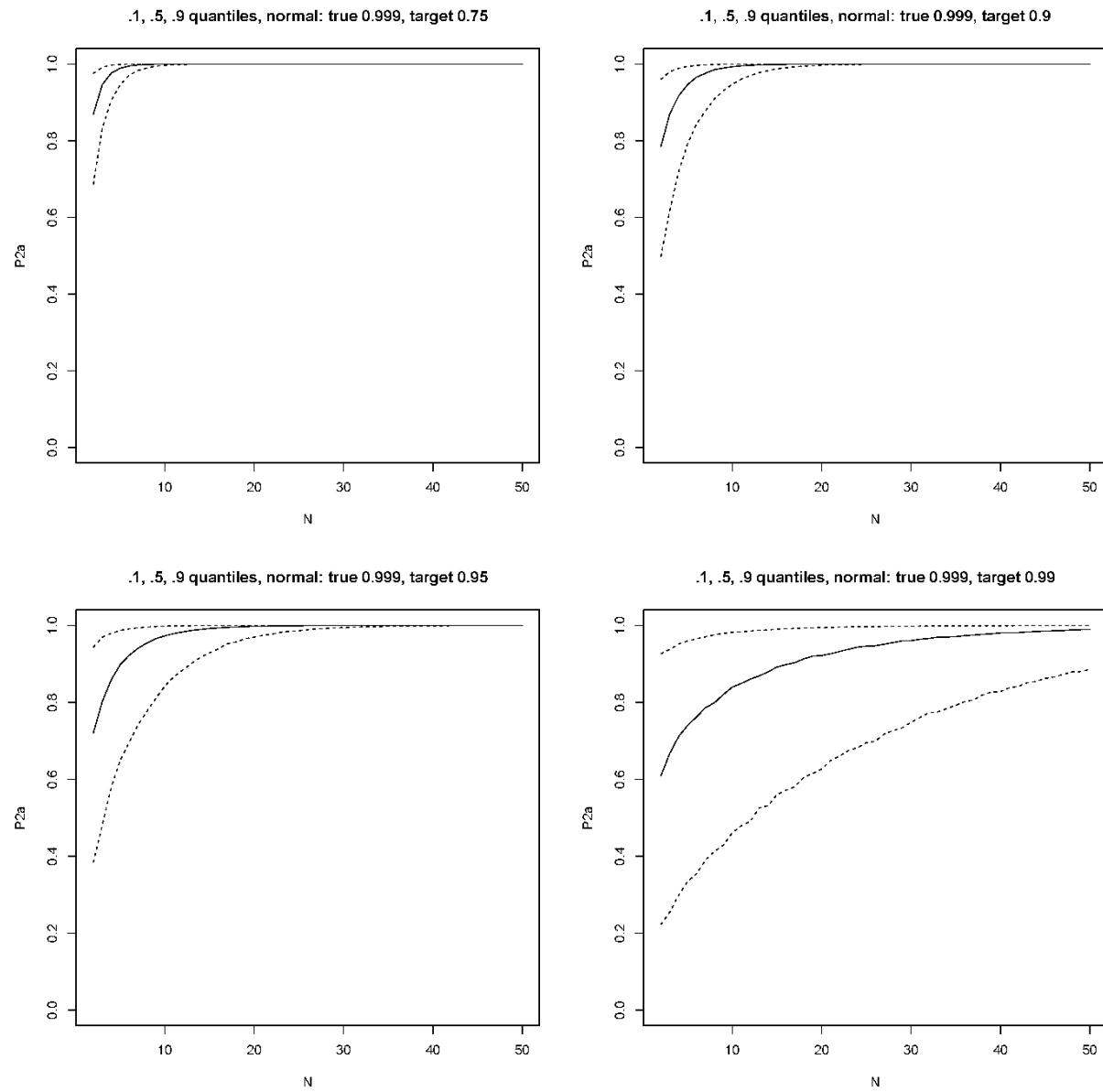
P2a Results, True=0.99, Varying Target



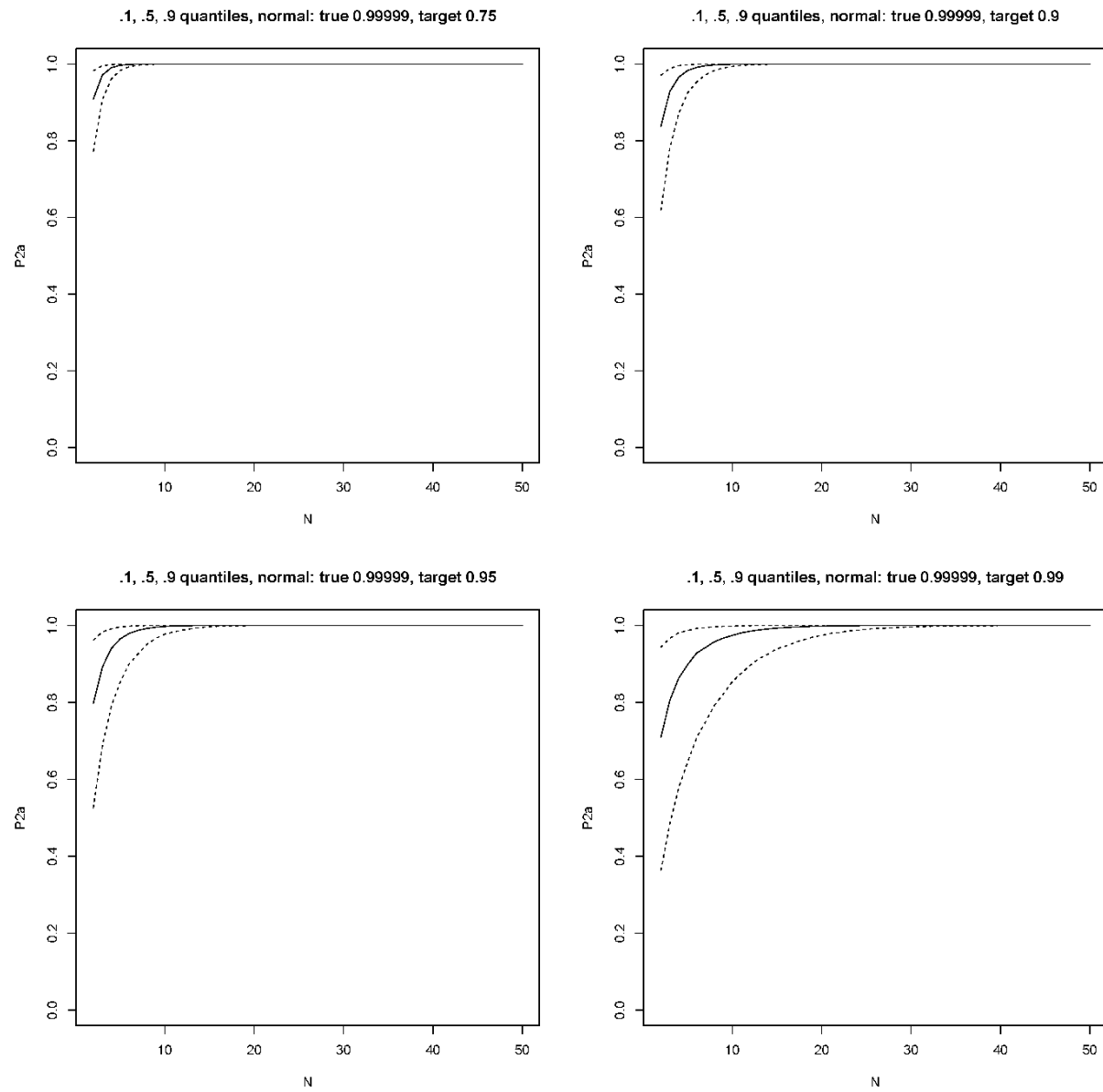
P2a Results, True=0.995, Varying Target



P2a Results, True=0.999, Varying Target

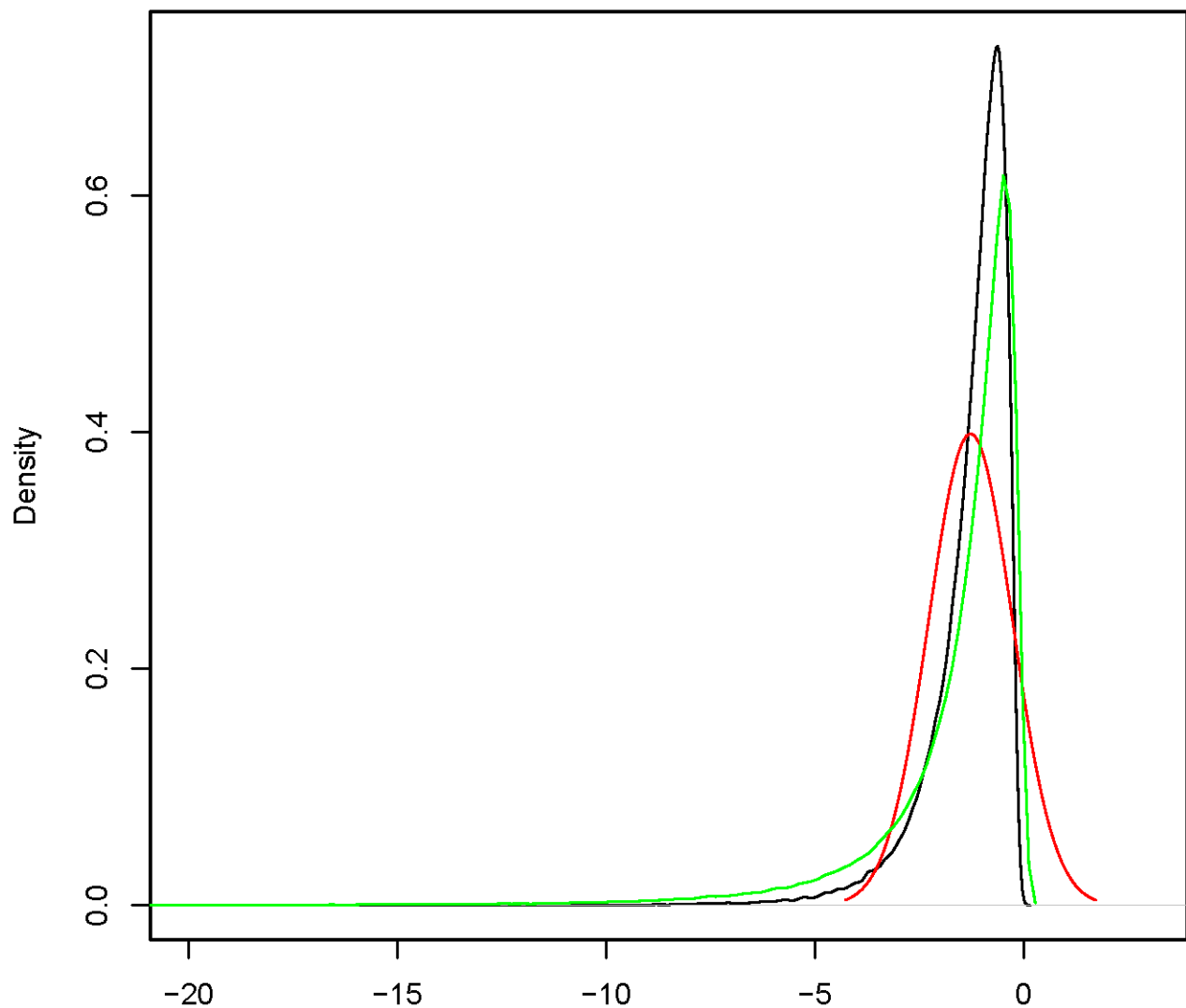


P2a Results, True=0.99999, Varying Target



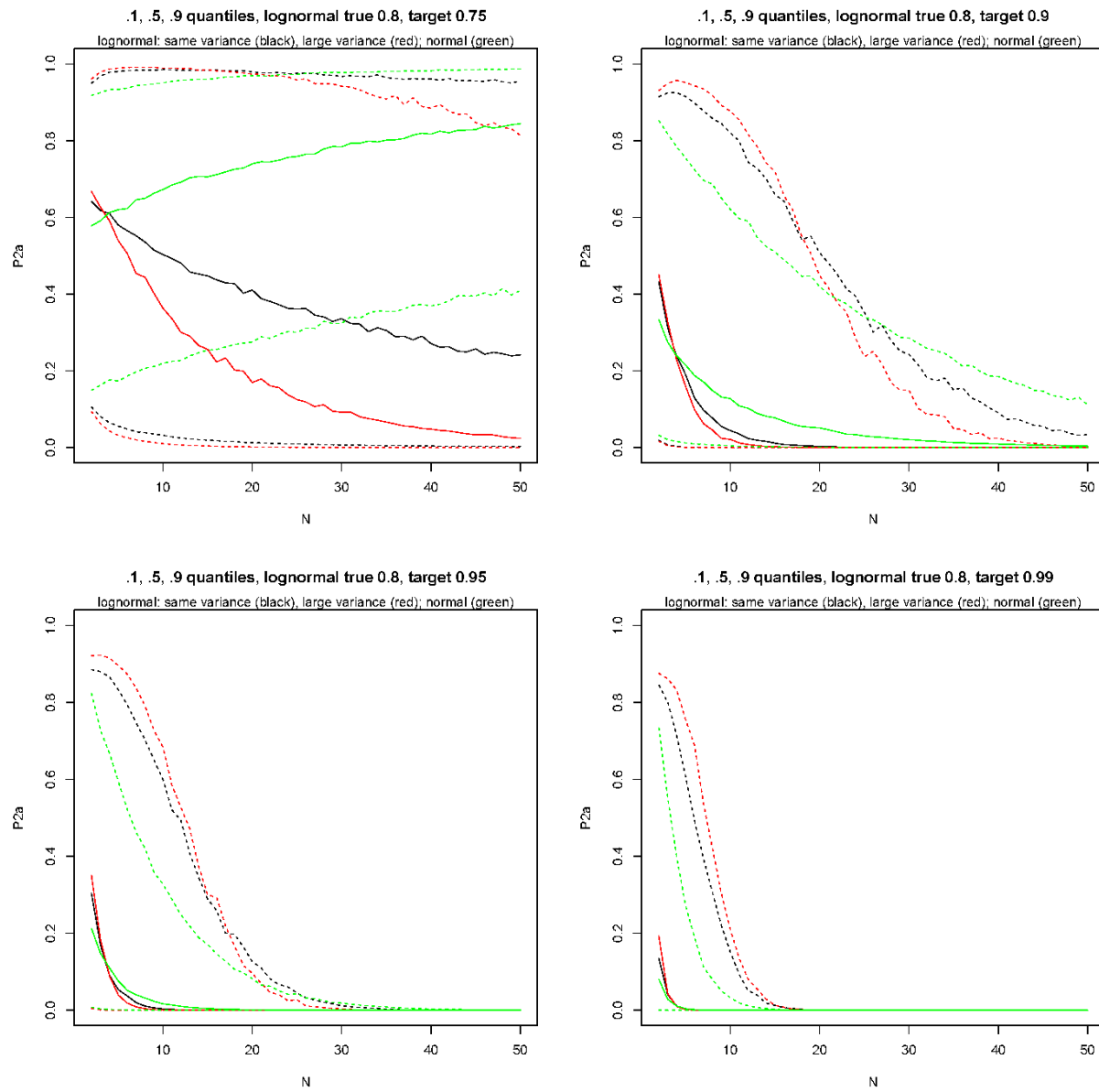
APPENDIX D: METRIC 2A SENSITIVITY STUDY GRAPHS, LOG-NORMAL DISTRIBUTION

LN same var (black), normal (red), LN larger var (green)



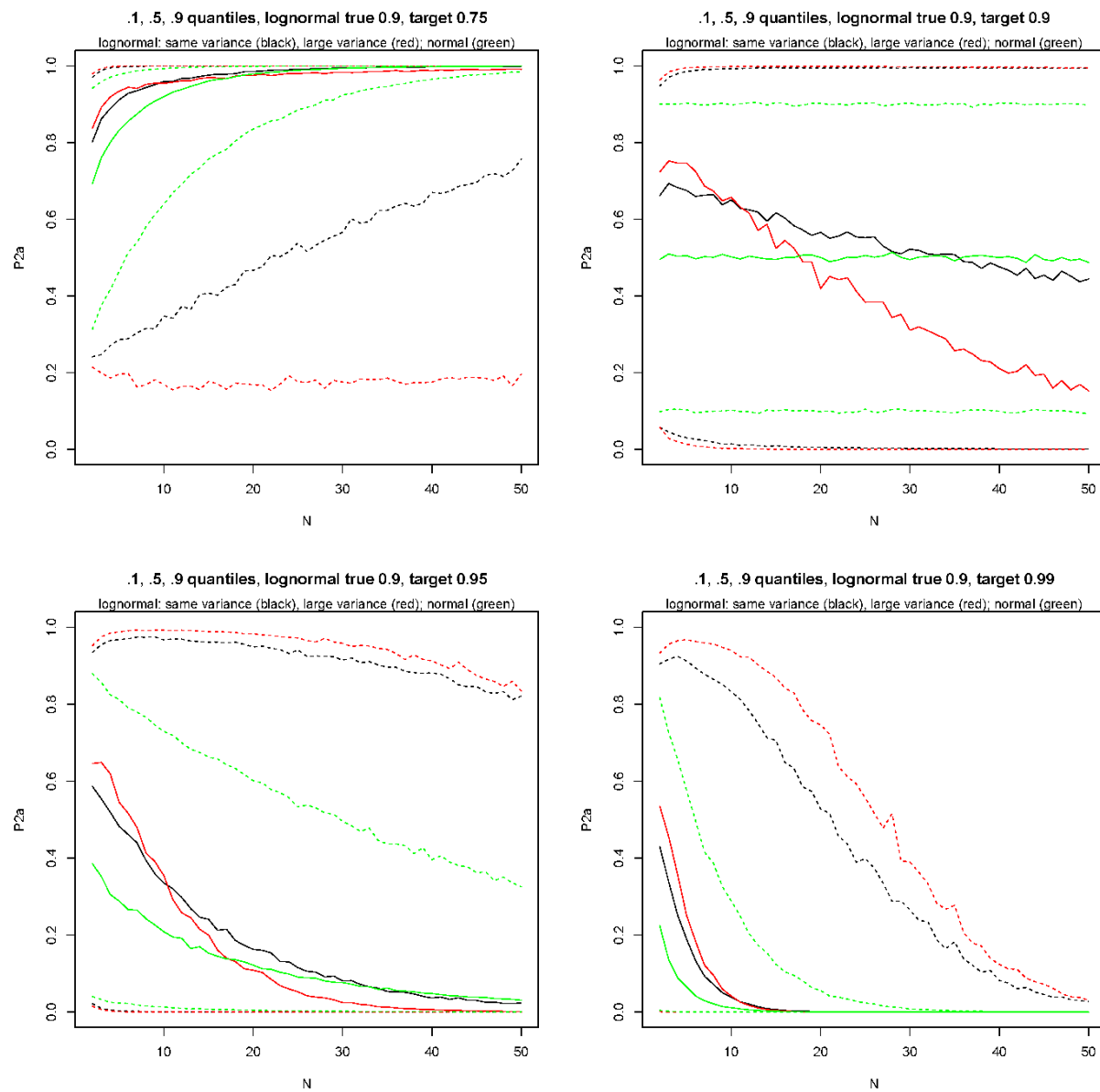
NOTE: On the graphs that follow, the **normal distribution** results are shown in green, the lognormal with the same variance in black, and the **lognormal with large variance** in red.

P2a Lognormal, True=0.8, Varying Target



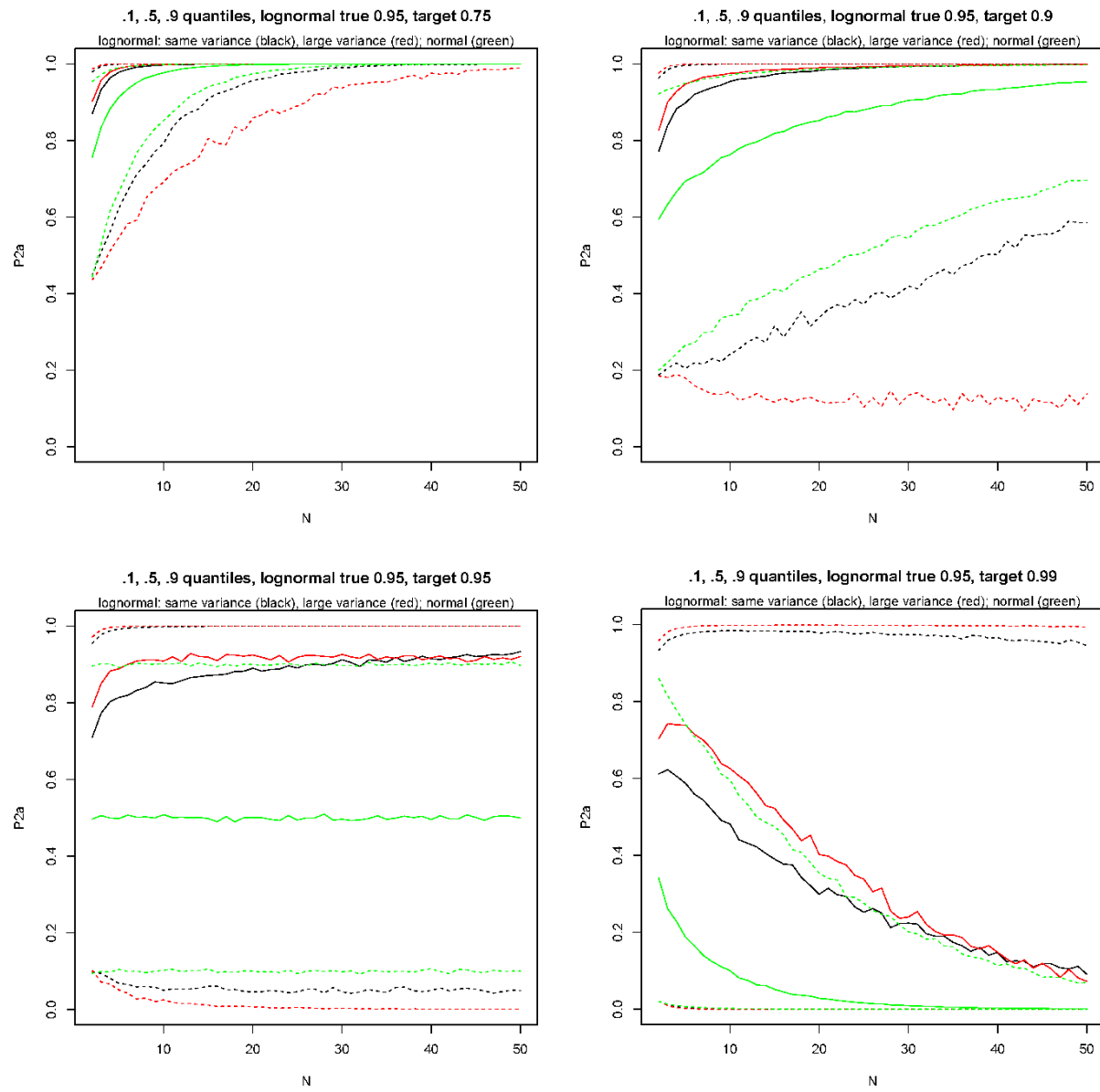
P2a

Lognormal, True=0.9, Varying Target

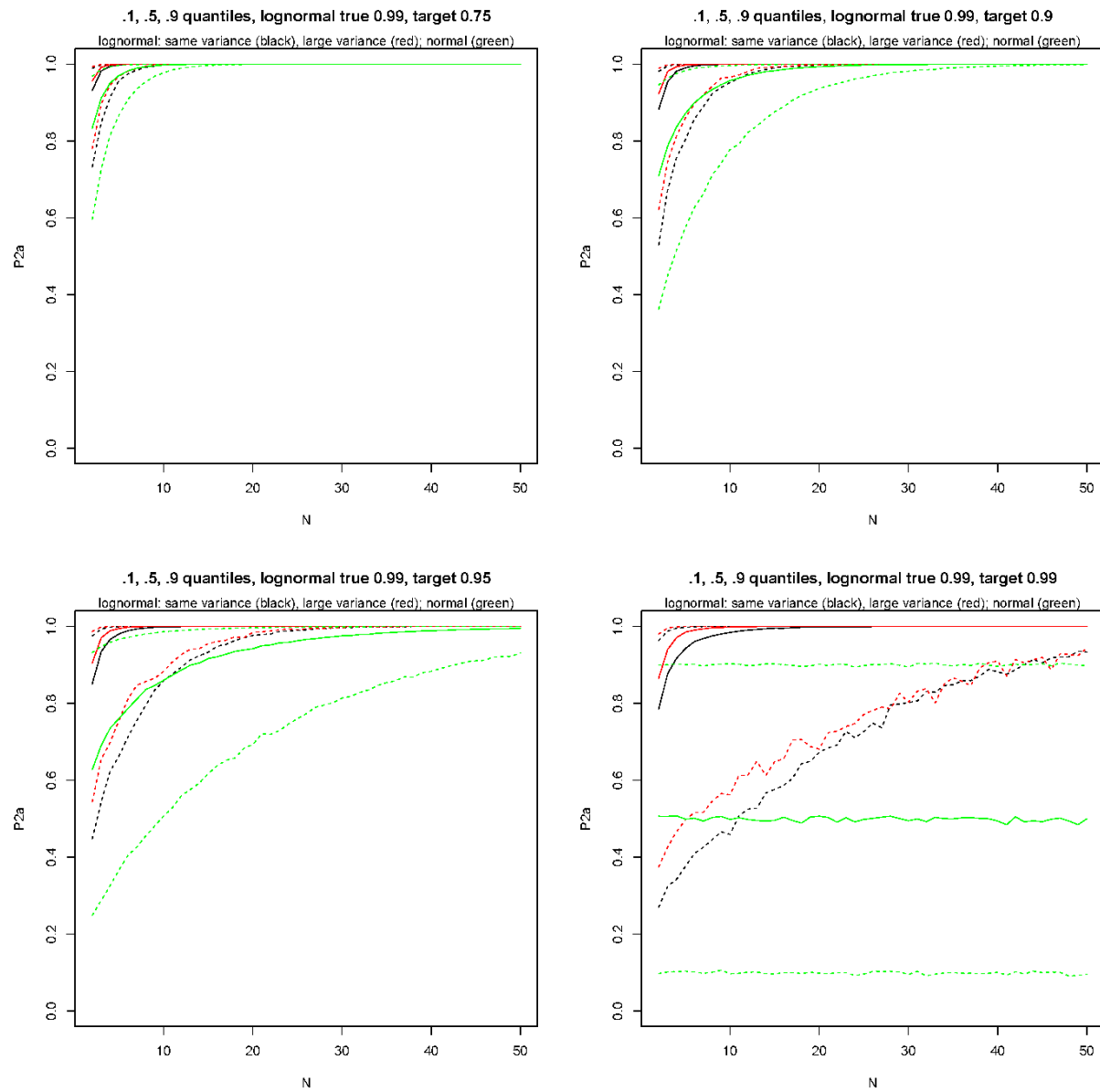


P_{2a}

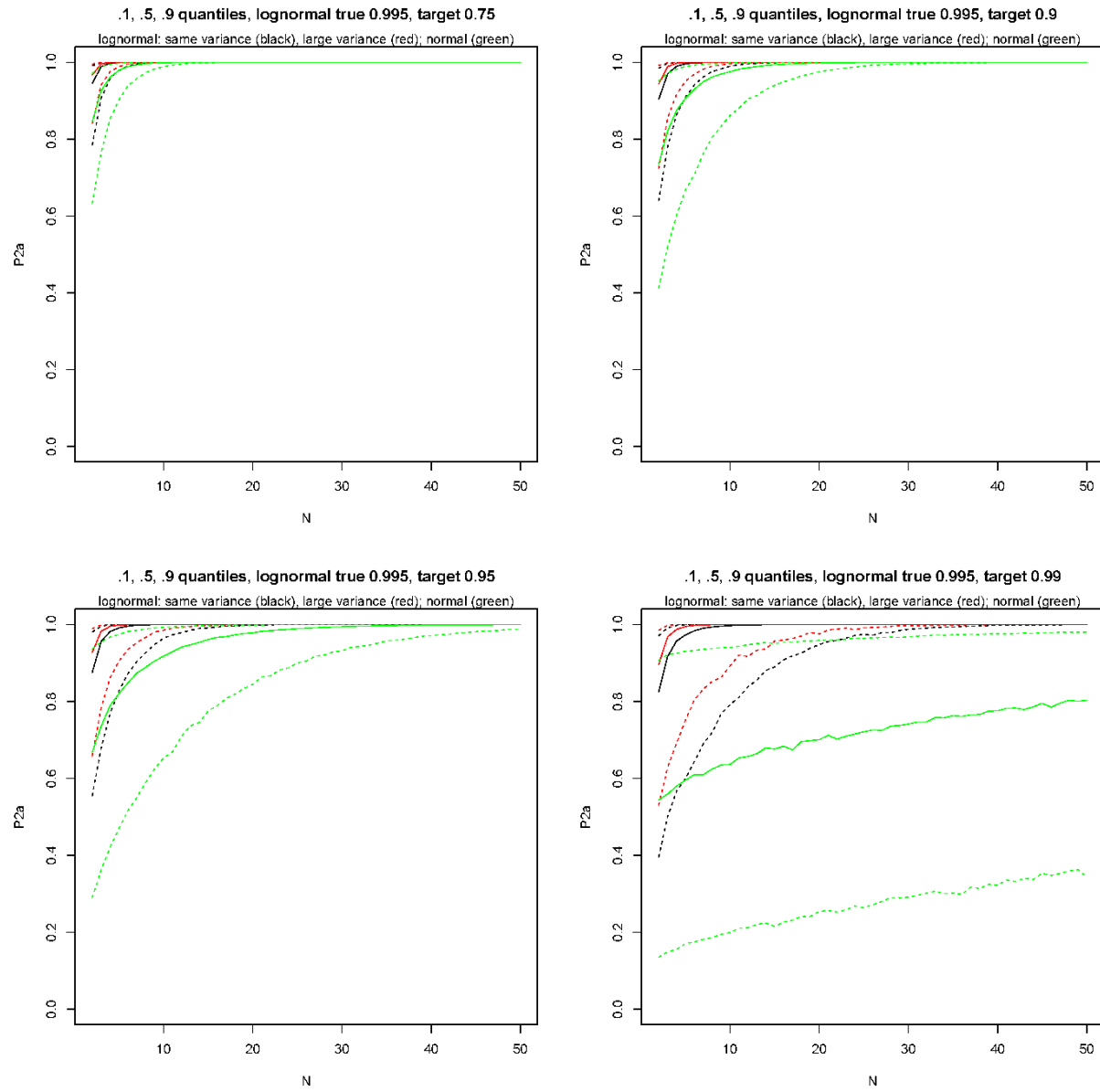
Lognormal, True=0.95, Varying Target



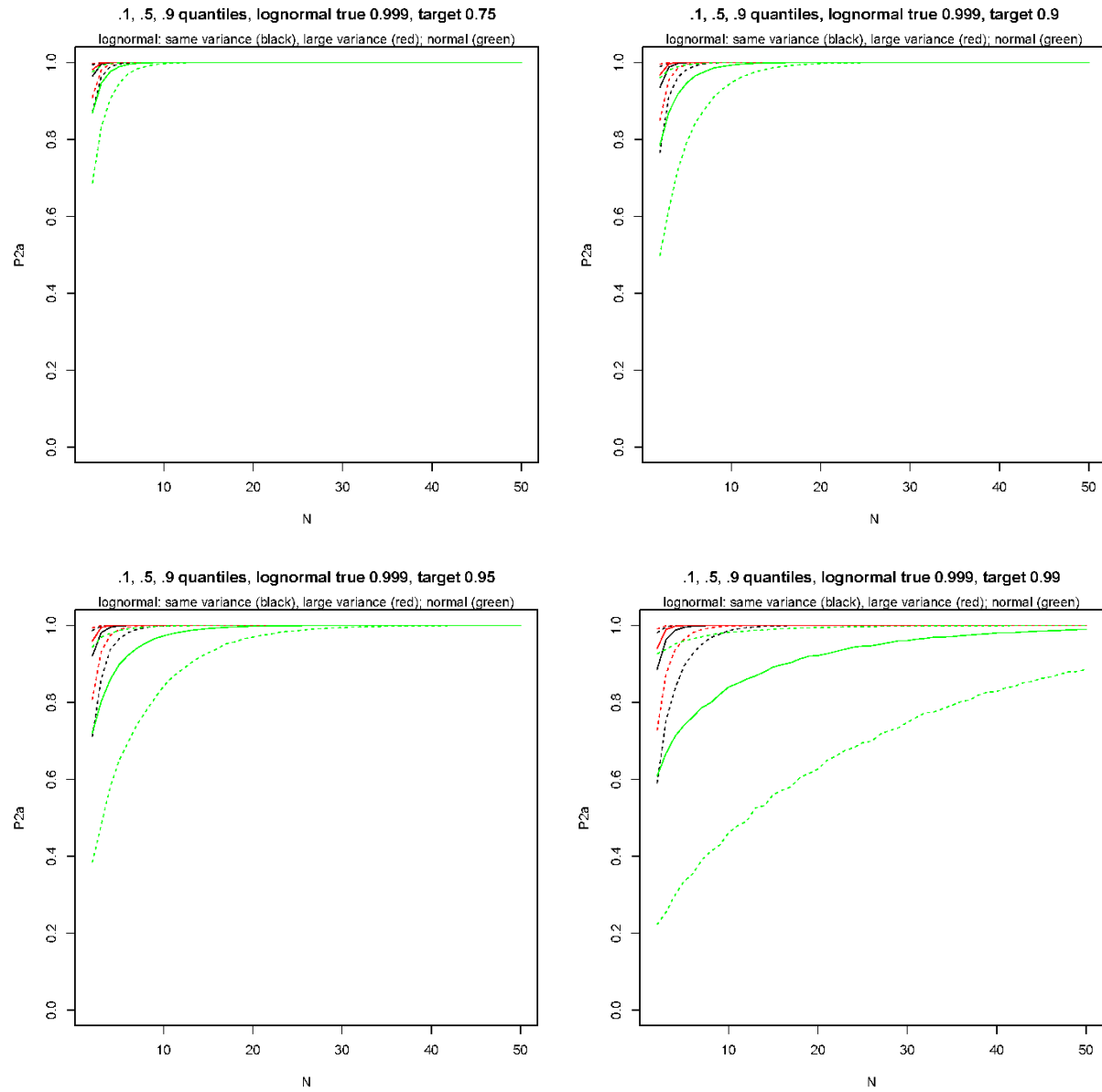
P2a Lognormal, True=0.99, Varying Target



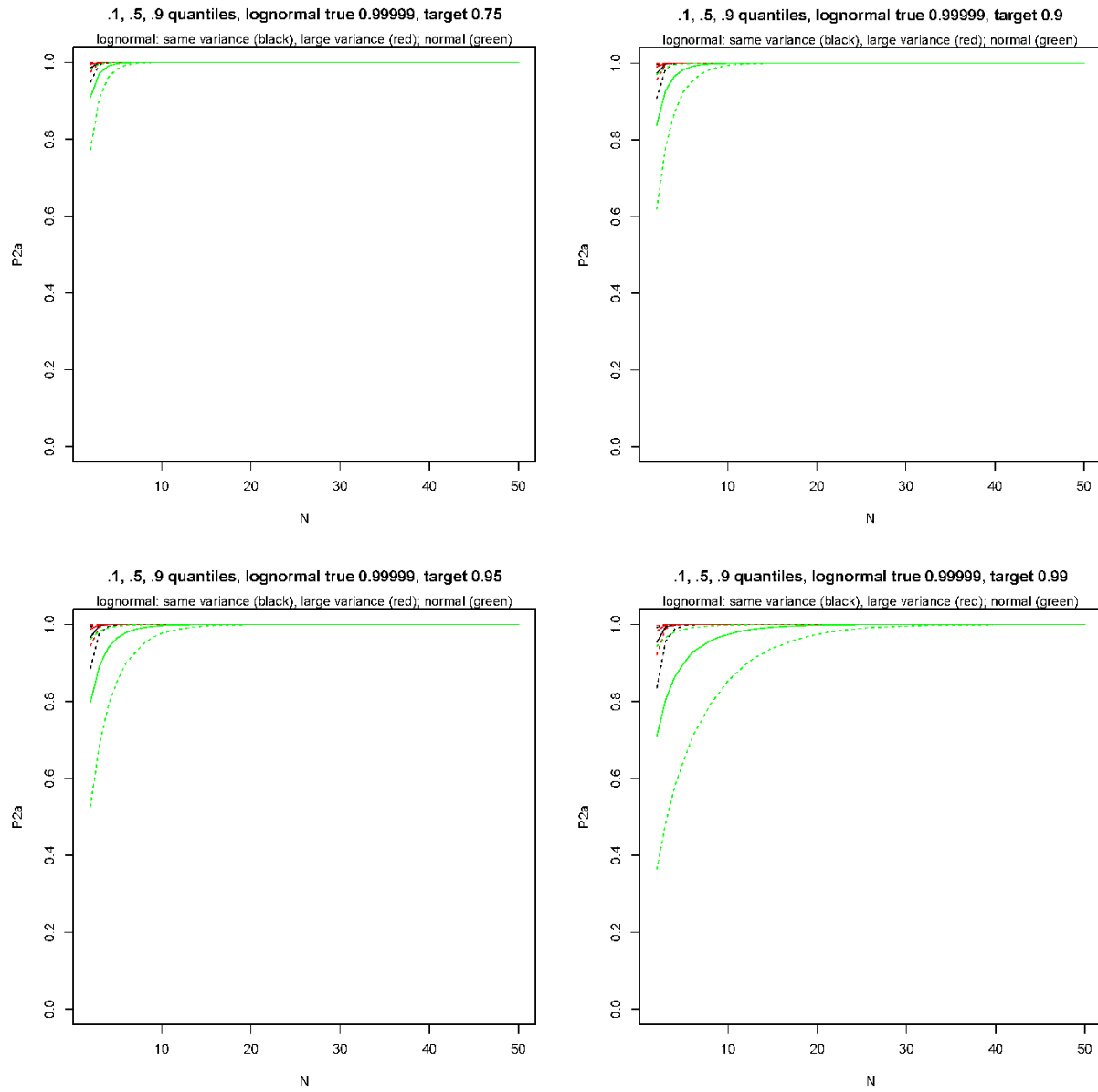
P2a Lognormal, True=0.995, Varying Target



P2a Lognormal, True=0.999, Varying Target

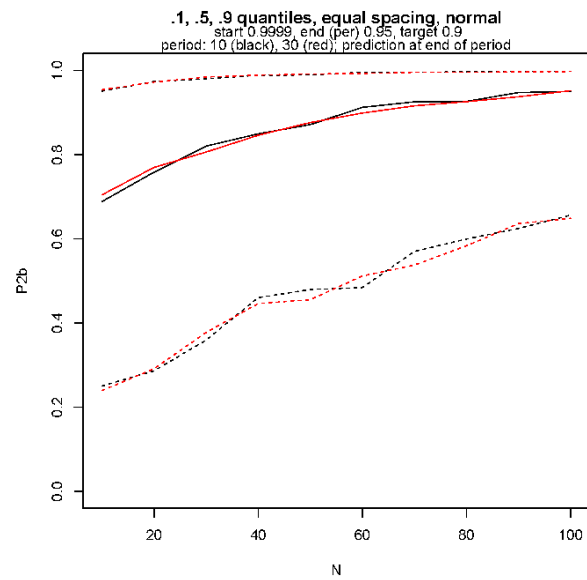
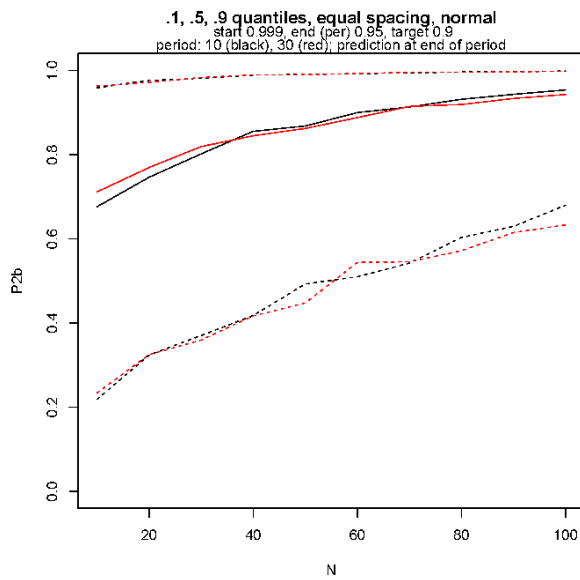
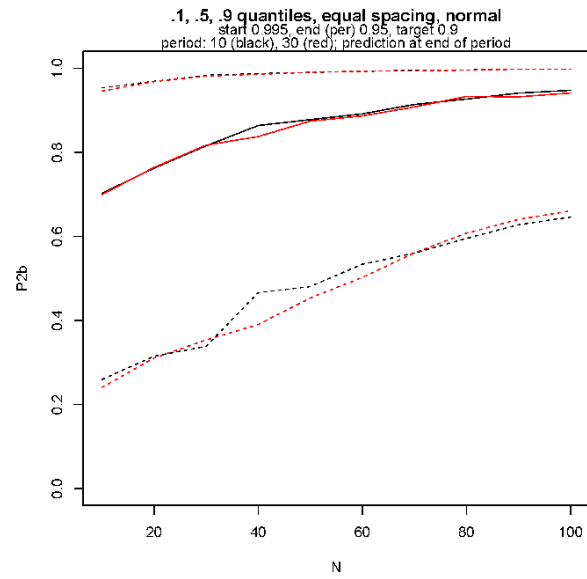
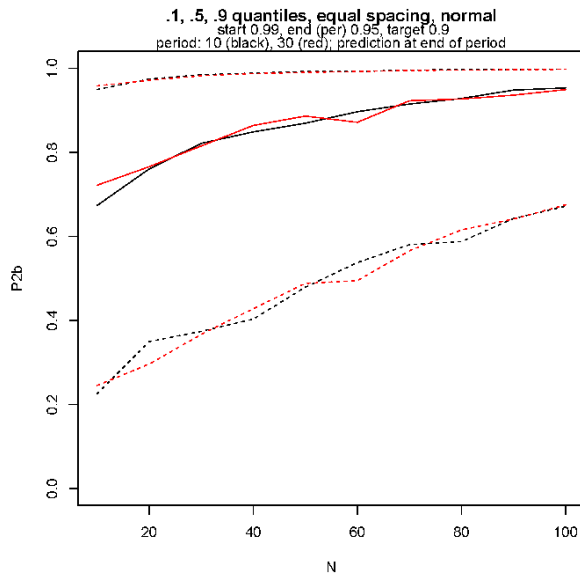


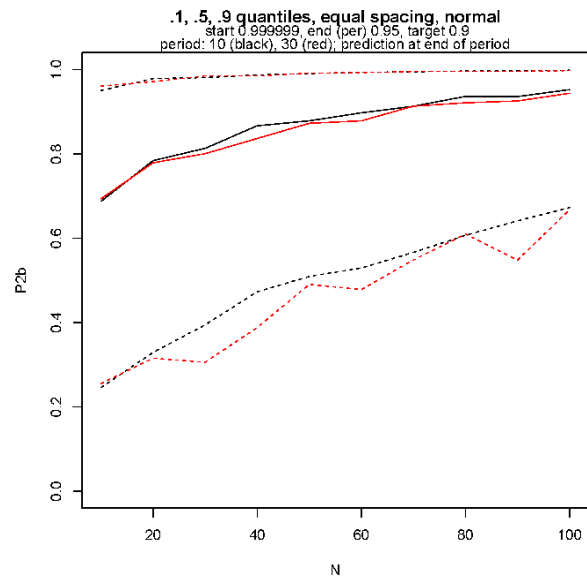
P2a Lognormal, True=0.99999, Varying Target



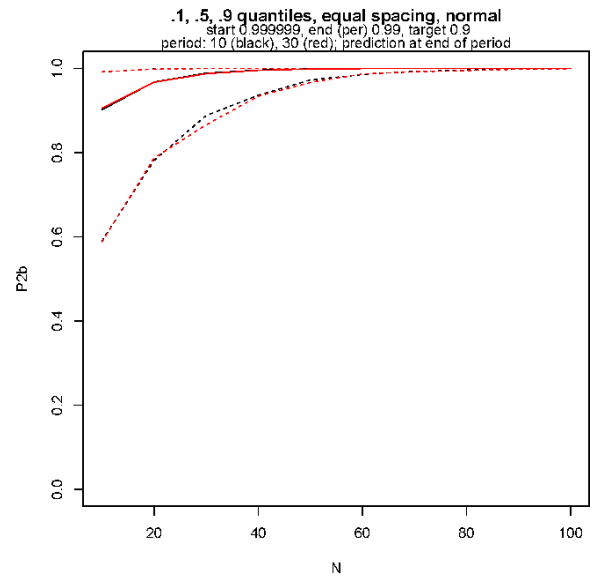
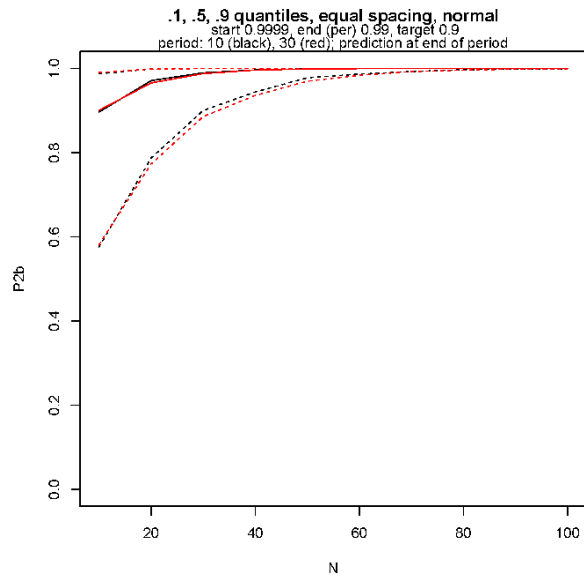
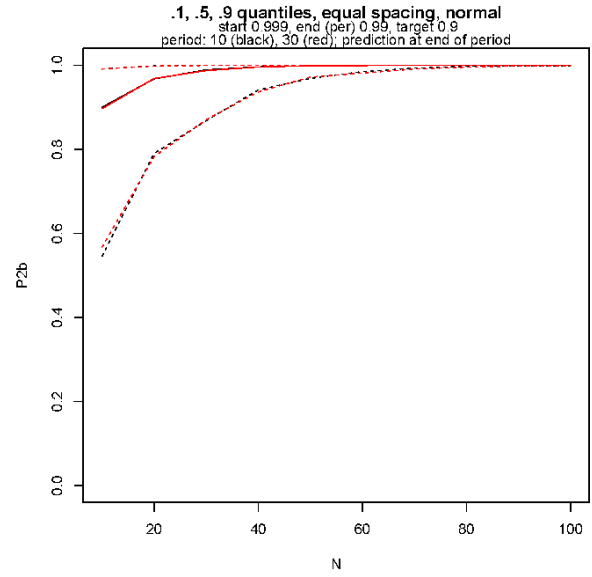
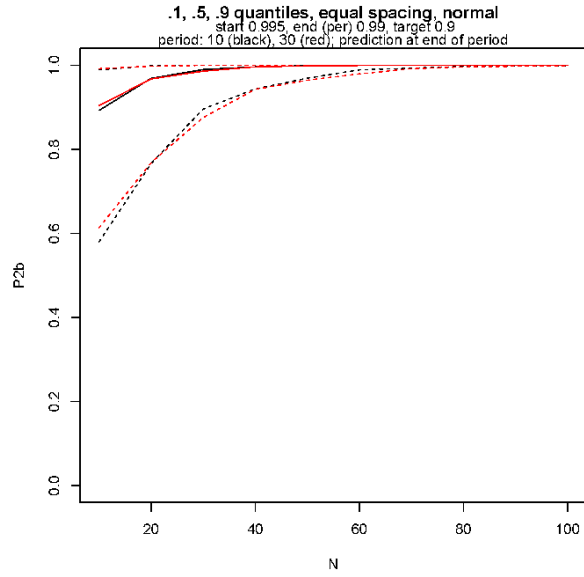
APPENDIX E: METRIC 2B SENSITIVITY STUDY GRAPHS

P2b Results, End=0.95, Target=0.9, Varying Start

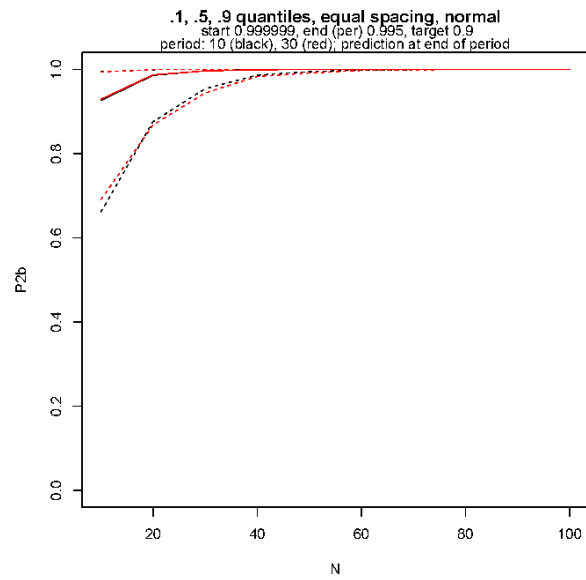
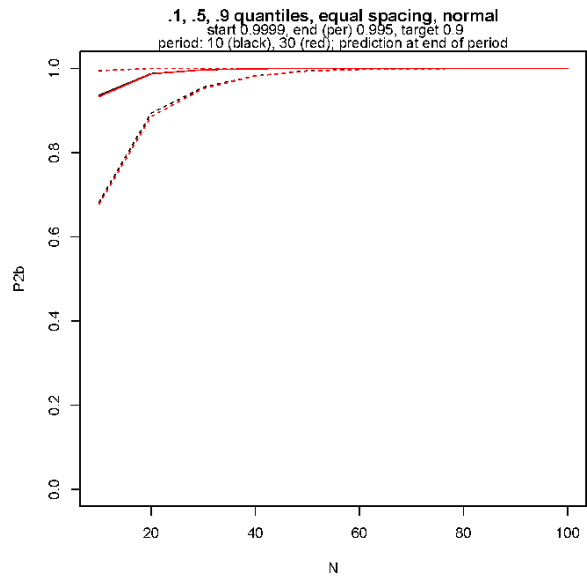
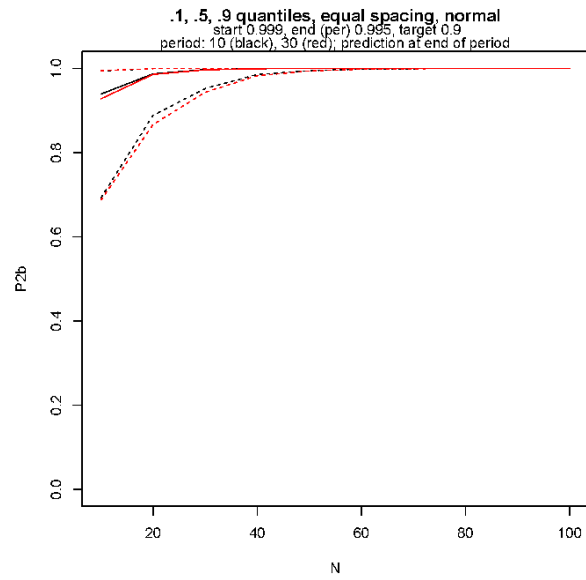




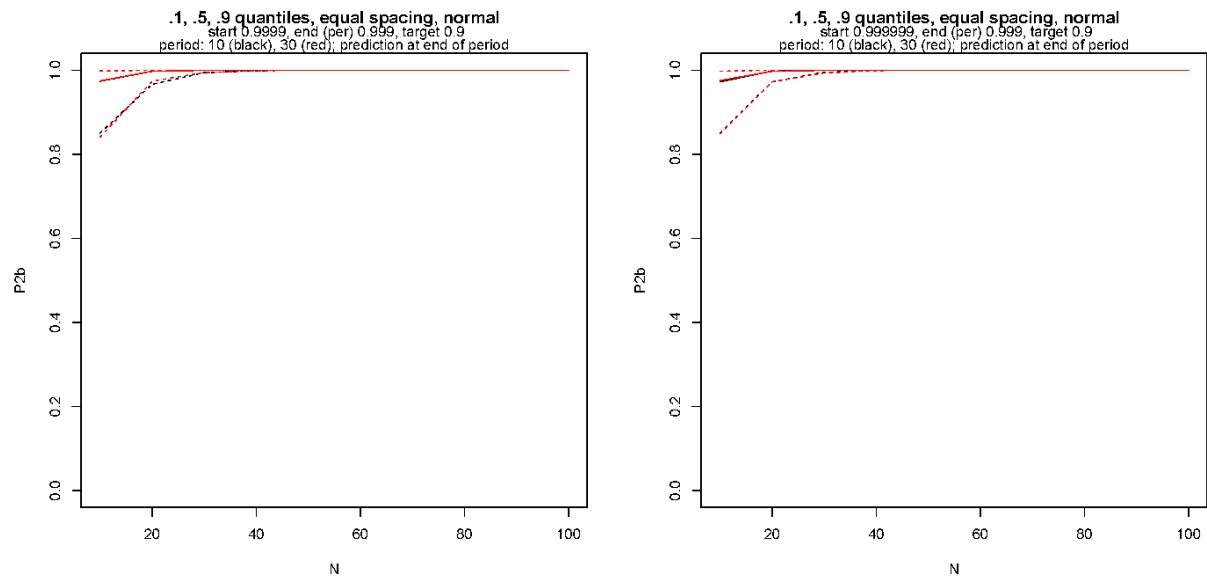
P2b Results, End=0.99, Target=0.9, Varying Start



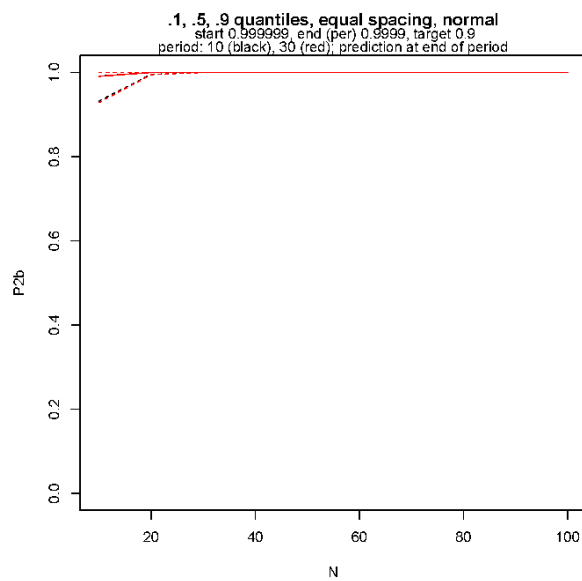
P2b Results, End=0.995, Target=0.9, Varying Start



P2b Results, End=0.999, Target=0.9, Varying Start

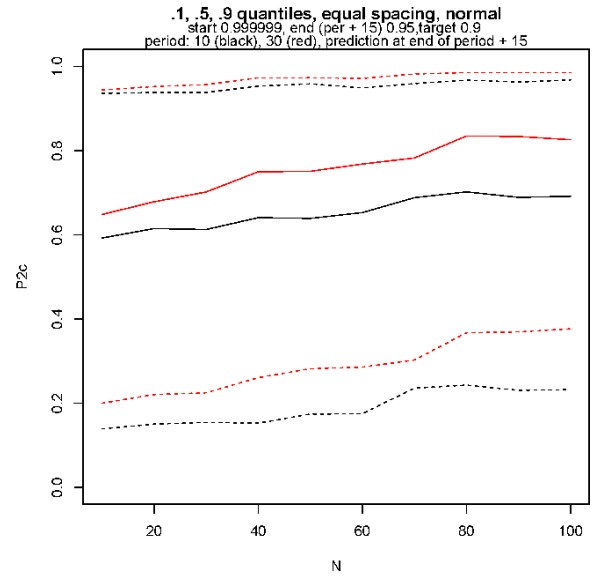
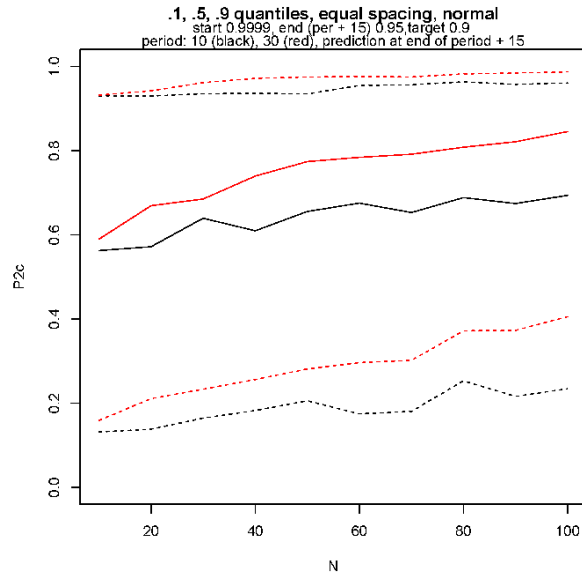
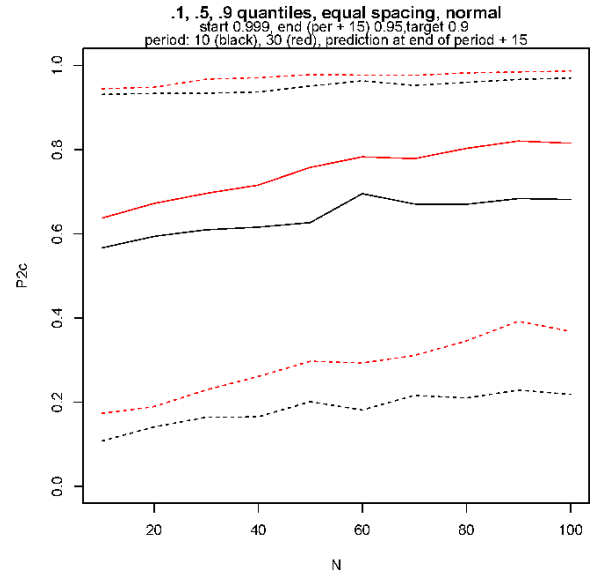
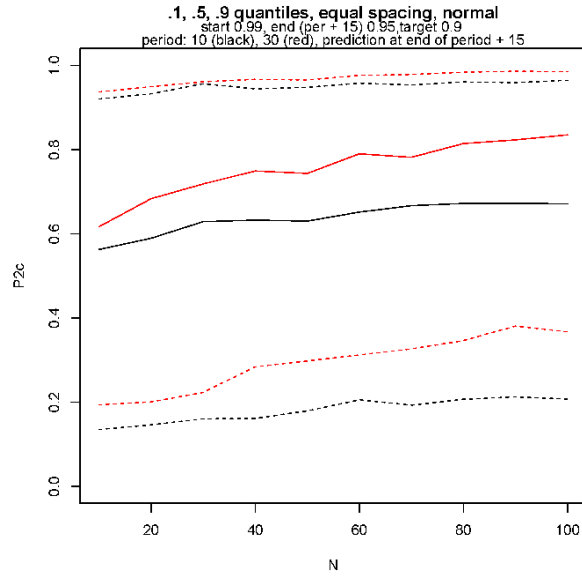


P2b Results, End=0.9999, Target=0.9, Varying Start

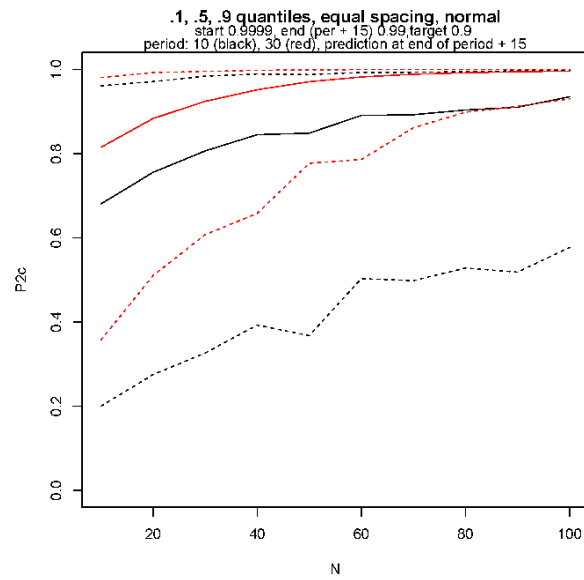
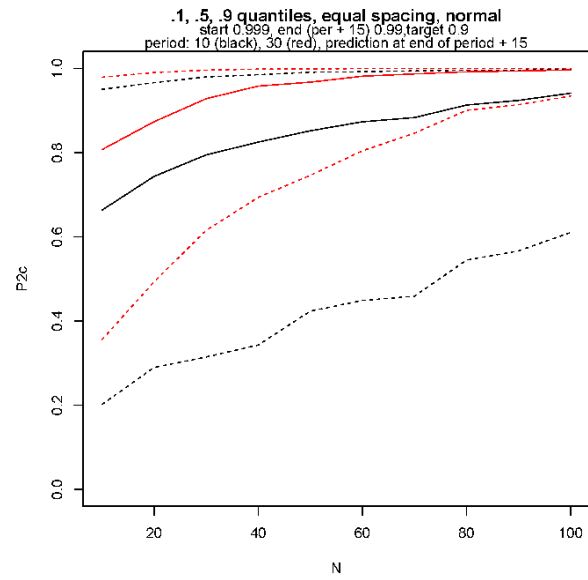
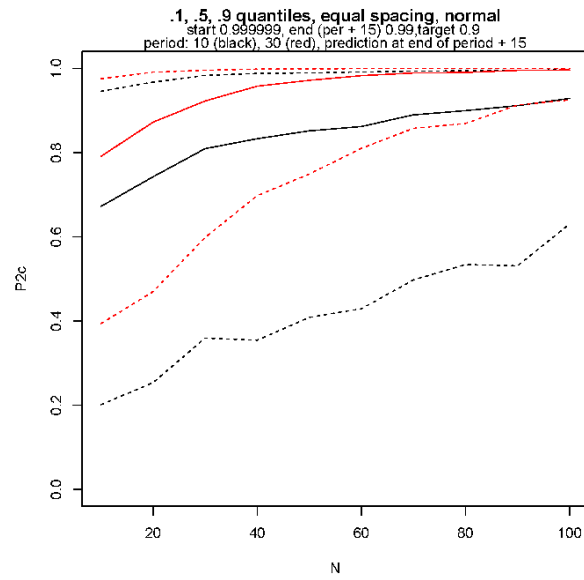


APPENDIX F: METRIC 2C SENSITIVITY STUDY GRAPHS

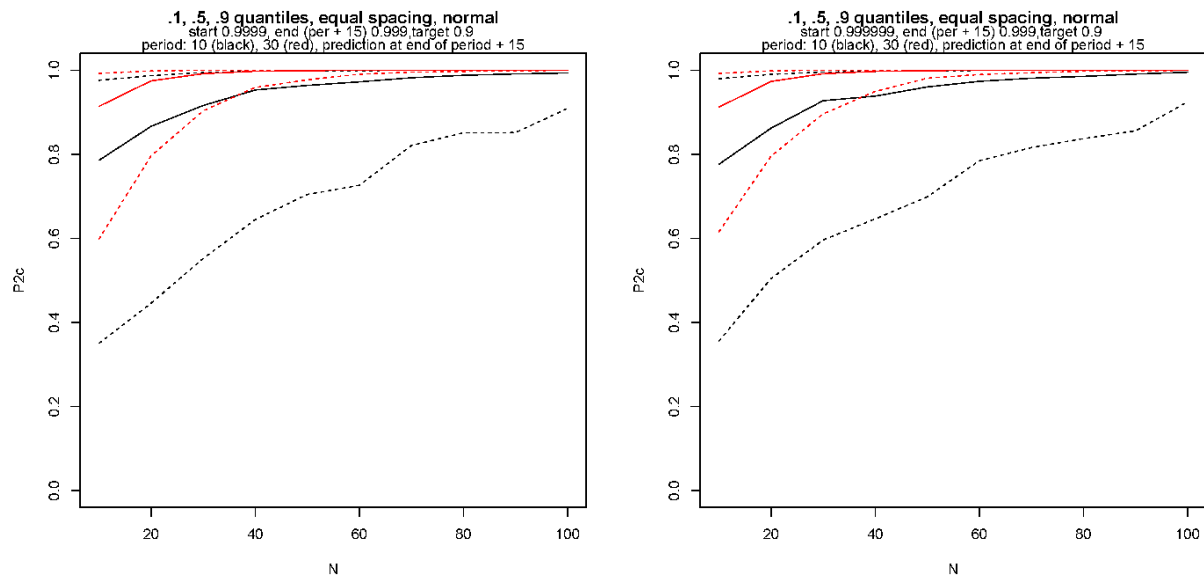
P2c Results, End=0.95, Target=0.9, Varying Start



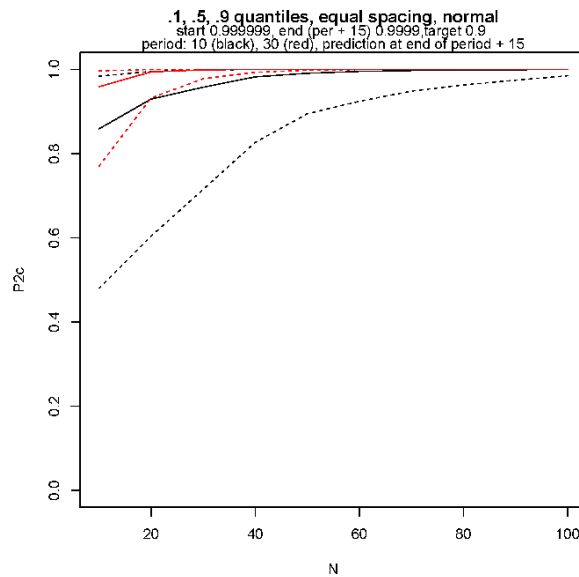
P2c Results, End=0.99, Target=0.9, Varying Start



P2c Results, End=0.999, Target=0.9, Varying Start

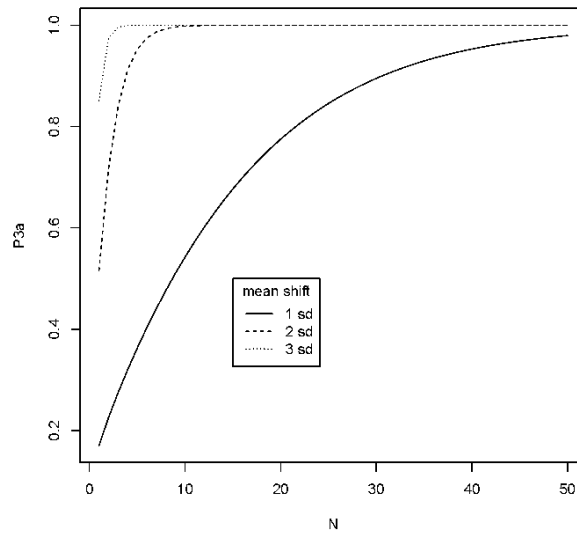


P2c Results, End=0.9999, Target=0.9, Varying Start

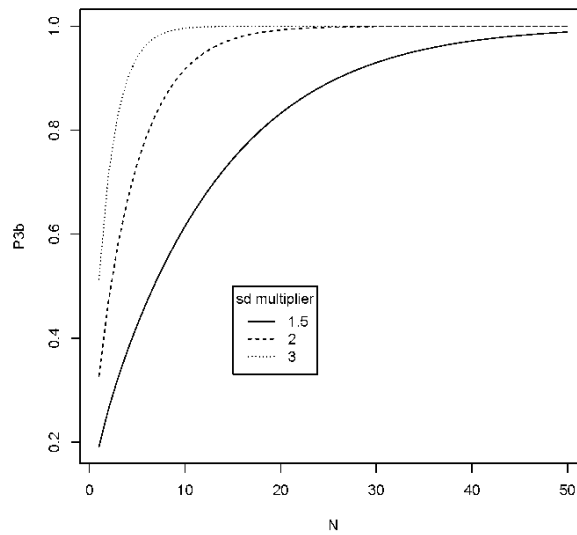


APPENDIX G: METRIC 3A/3B SENSITIVITY STUDY GRAPHS

P3a Results, Varying Mean Shift

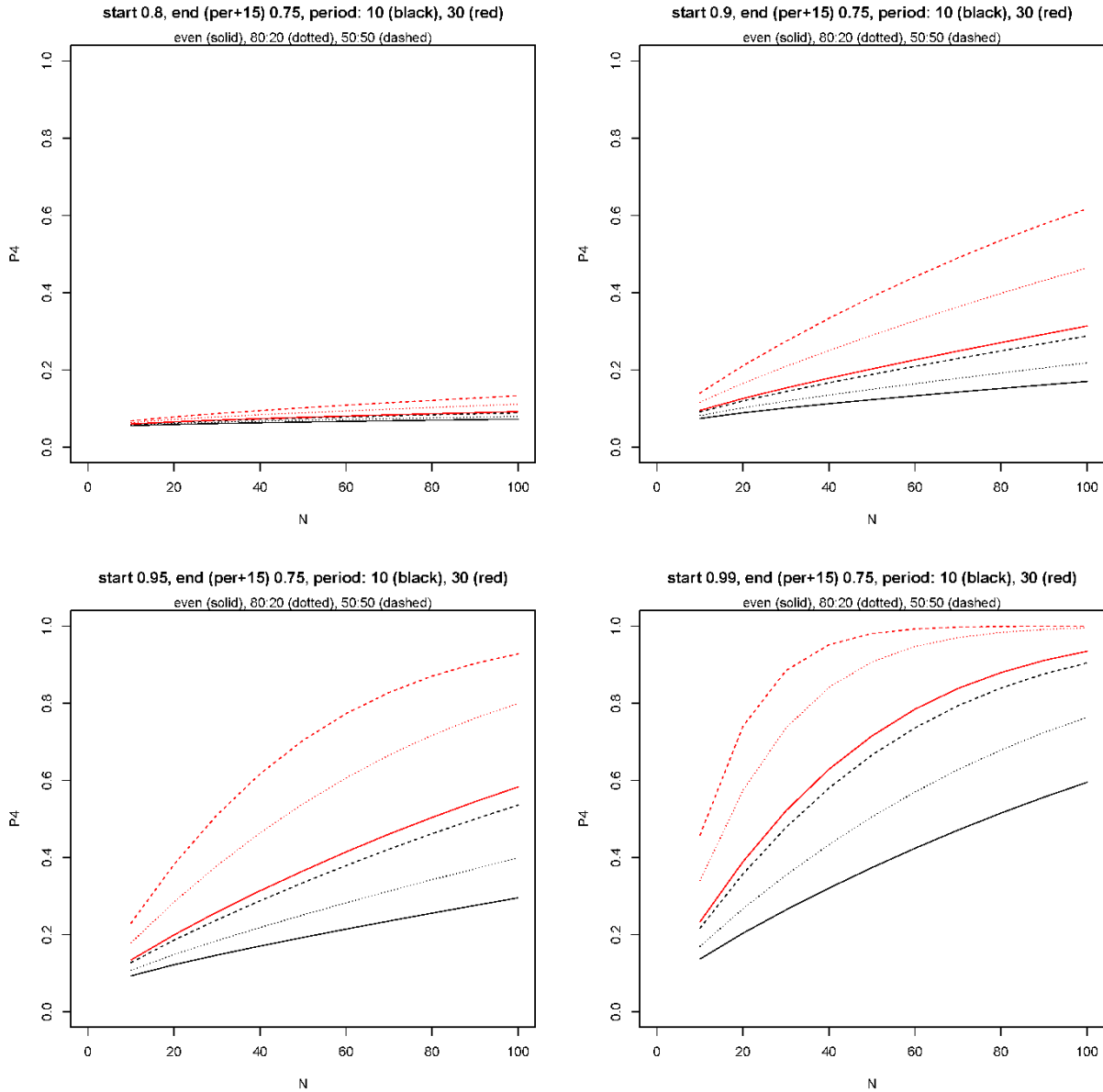


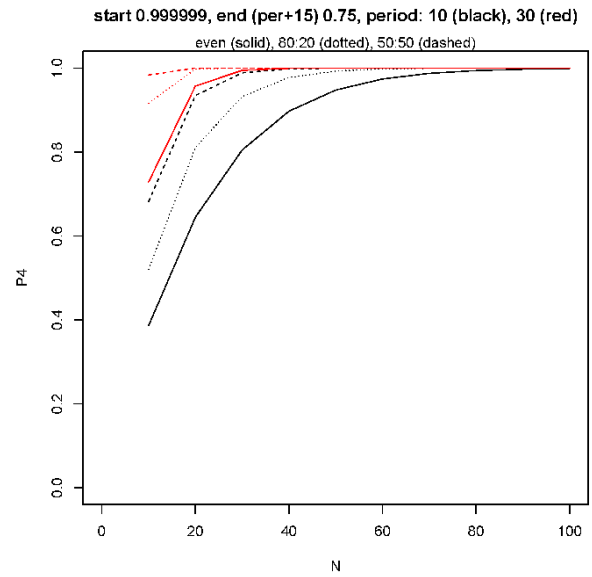
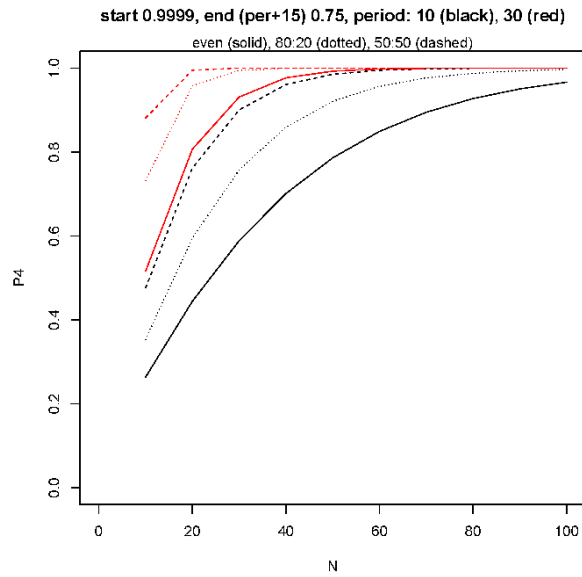
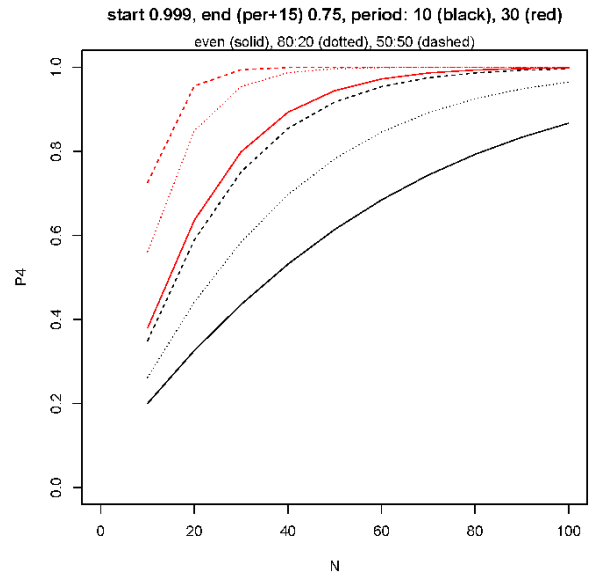
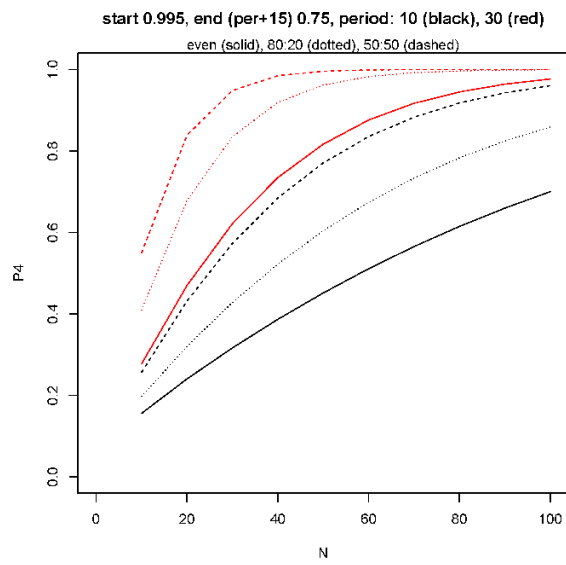
P3b Results, Varying Standard Deviation Multiplier



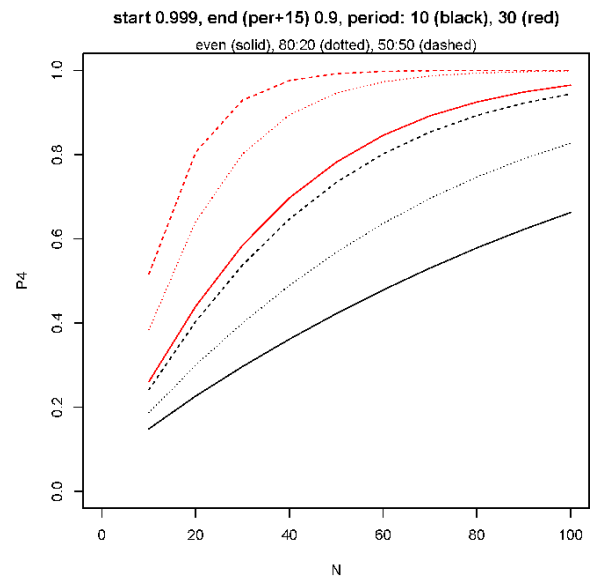
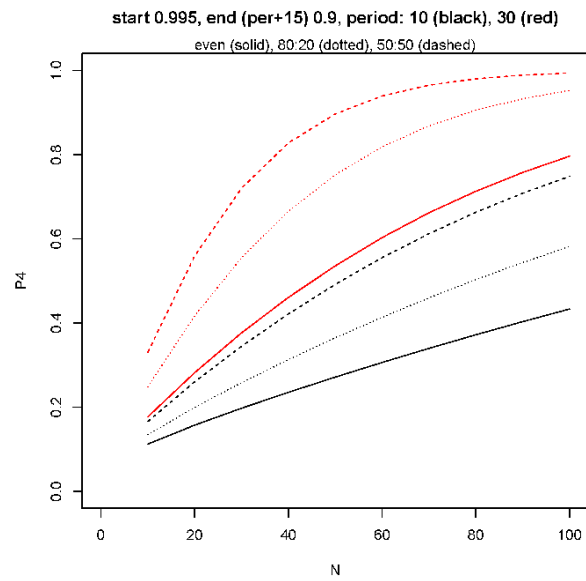
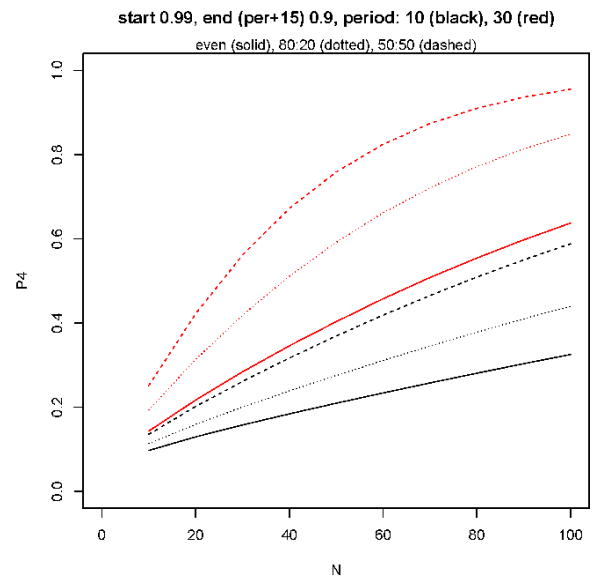
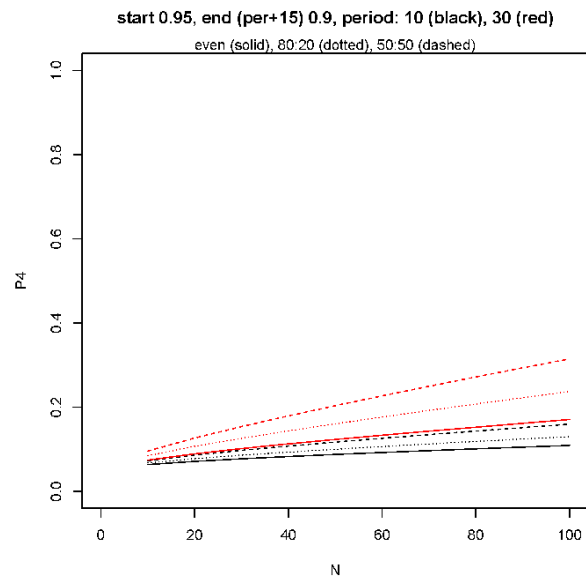
APPENDIX H: METRIC 4 SENSITIVITY STUDY GRAPHS, LINEAR TREND

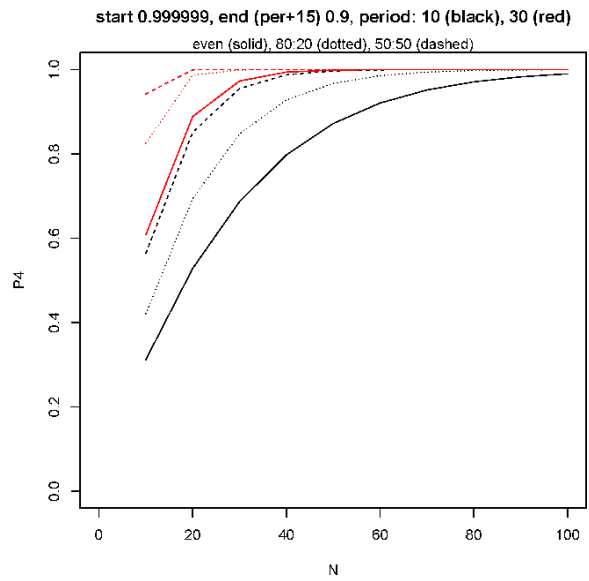
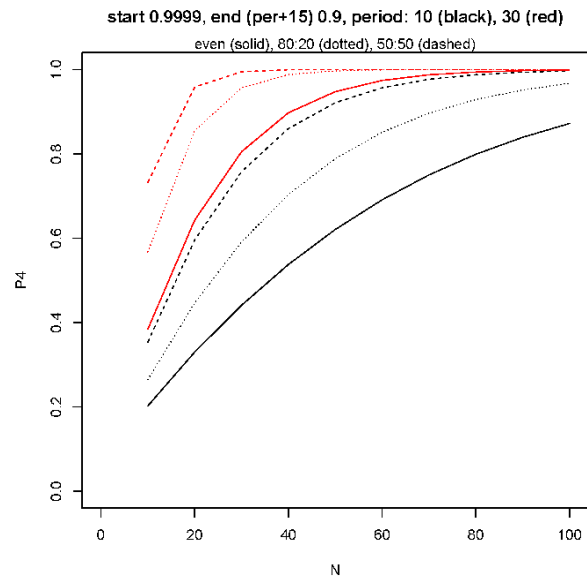
P4 Results, End=0.75, Varying Start



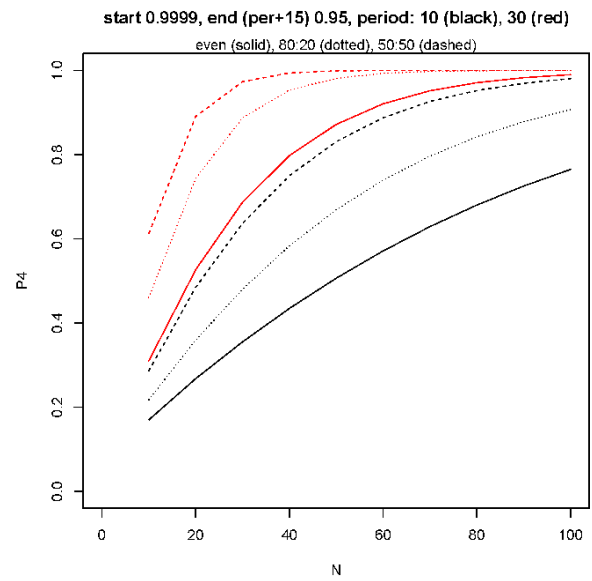
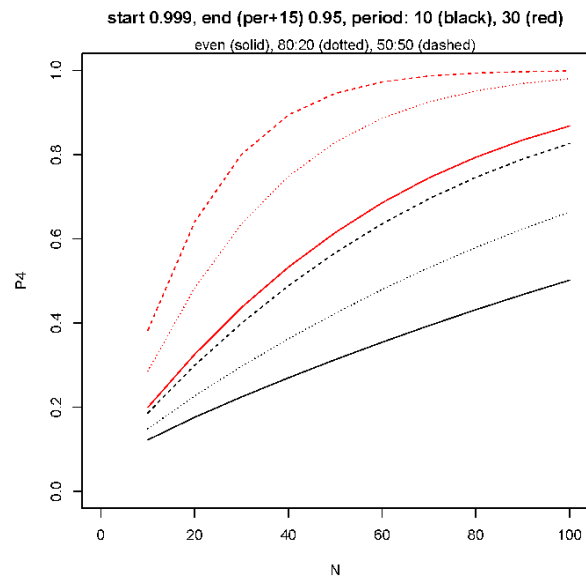
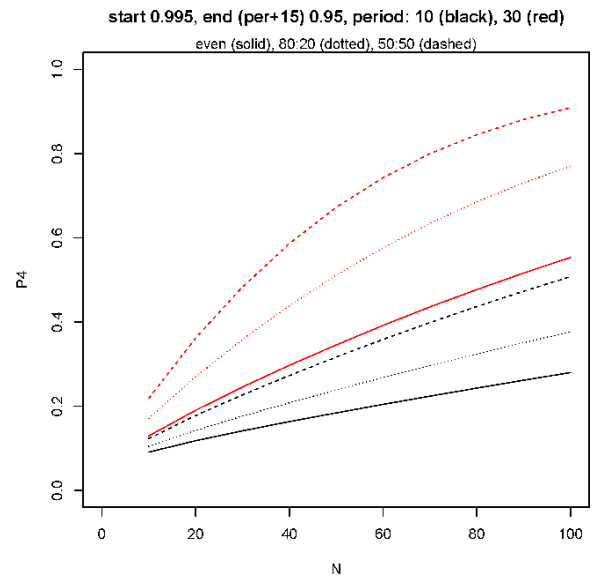
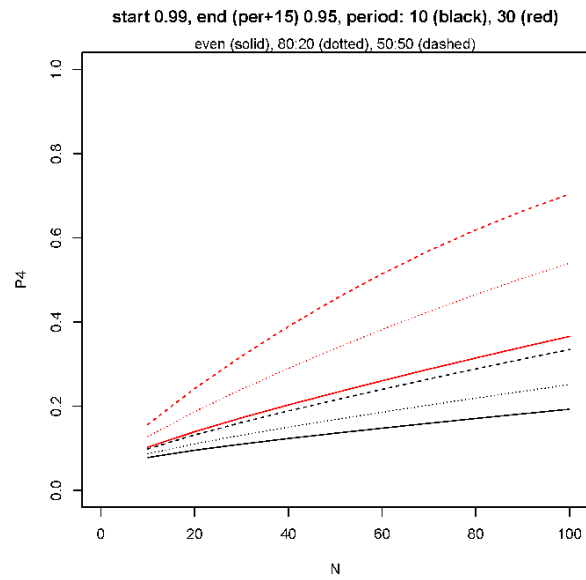


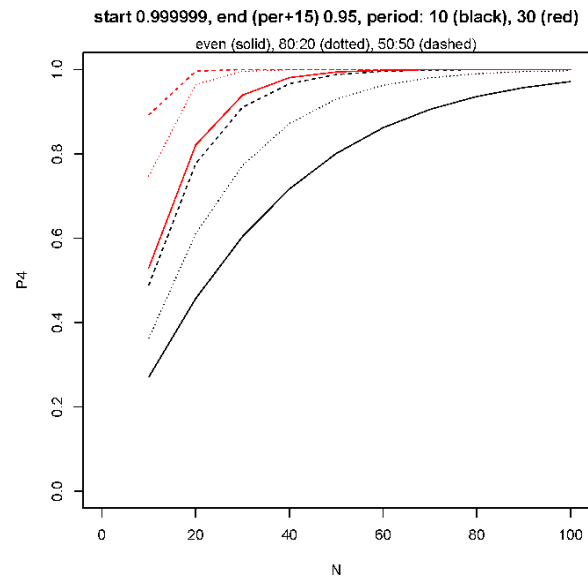
P4 Results, End=0.9, Varying Start



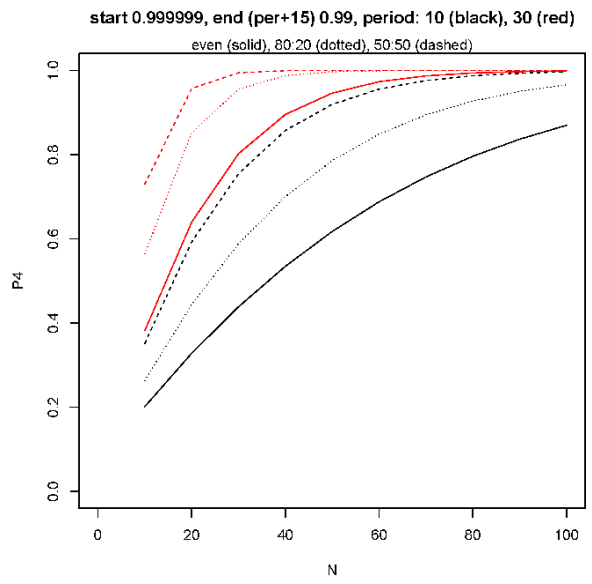
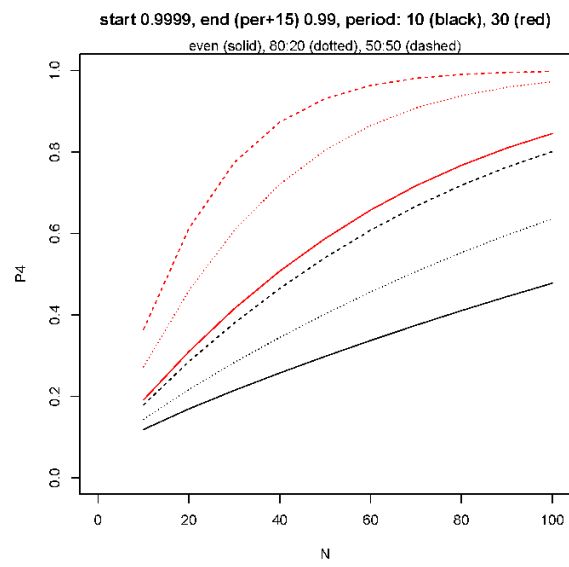
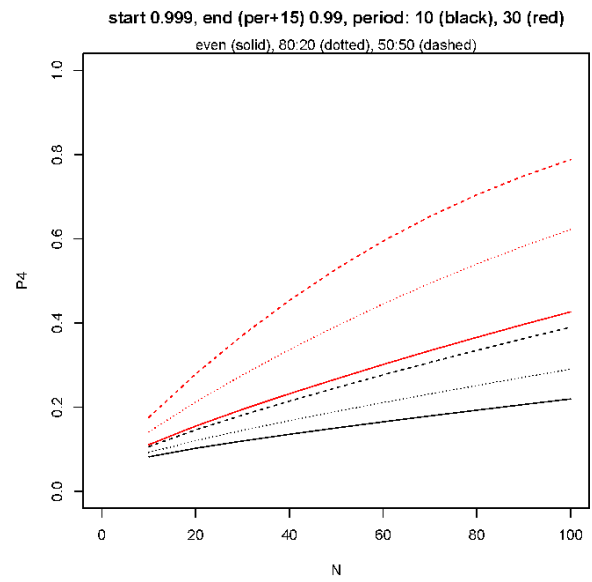
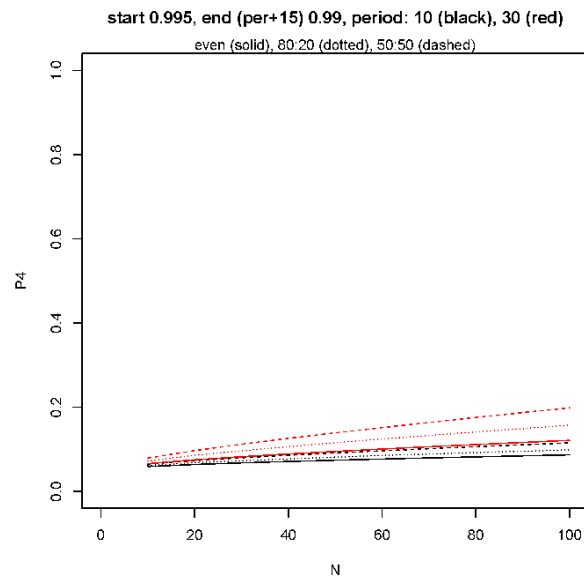


P4 Results, End=0.95, Varying Start

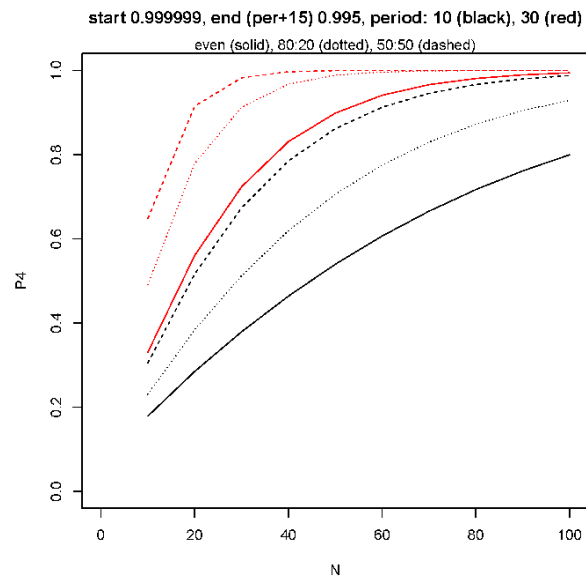
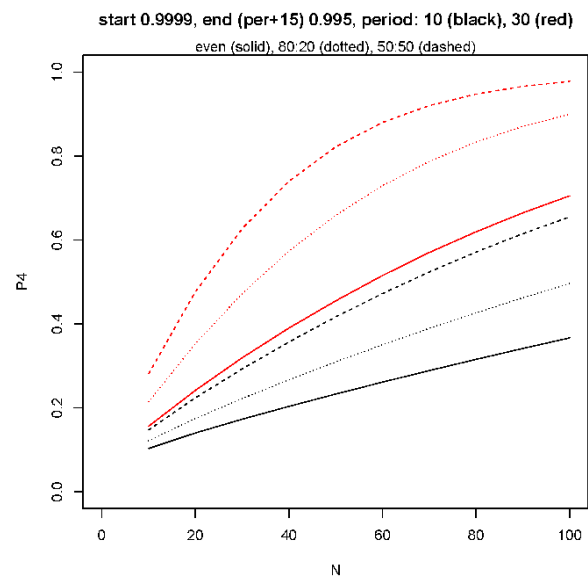
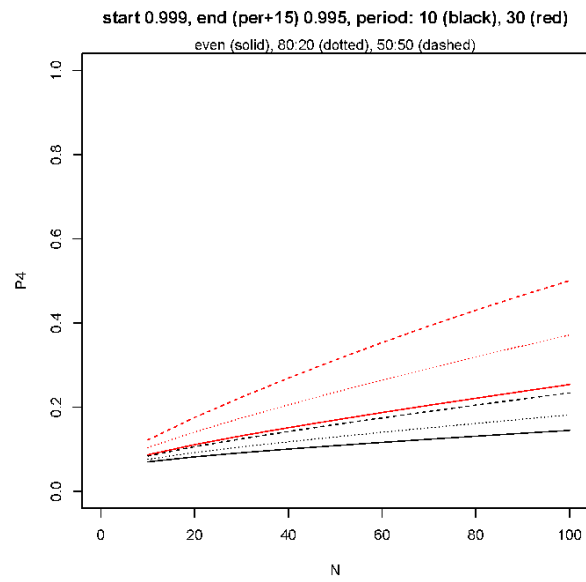




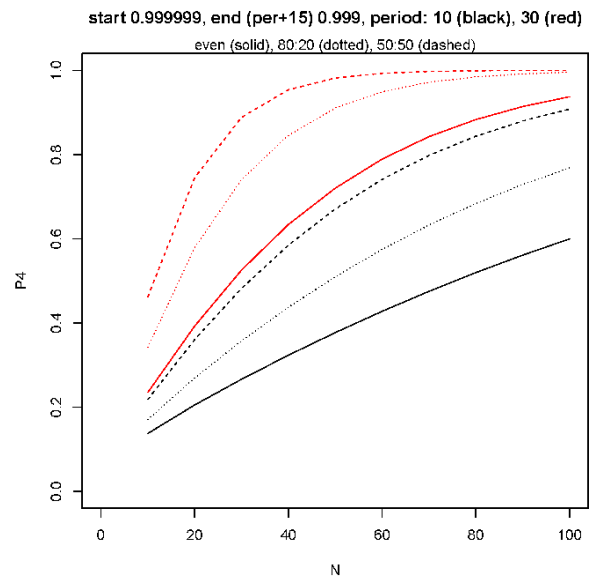
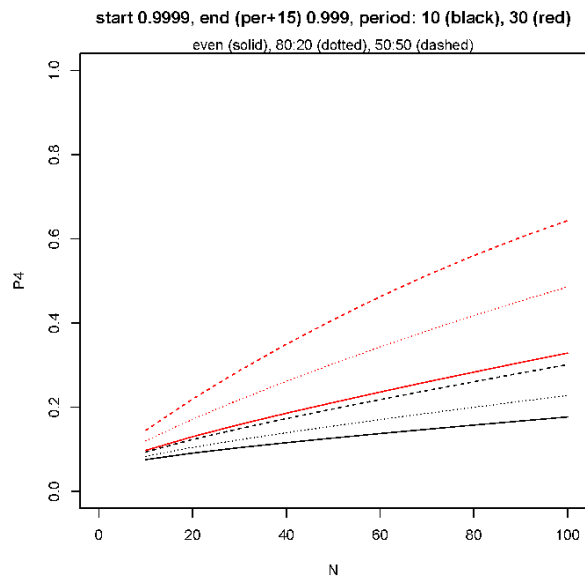
P4 Results, End=0.99, Varying Start



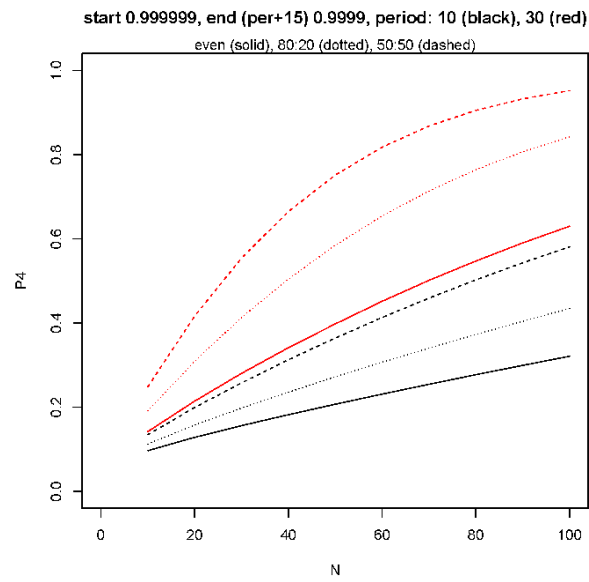
P4 Results, End=0.995, Varying Start



P4 Results, End=0.999, Varying Start

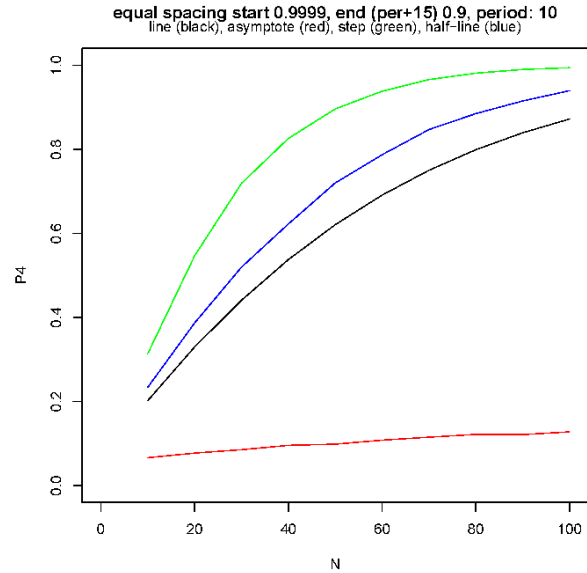
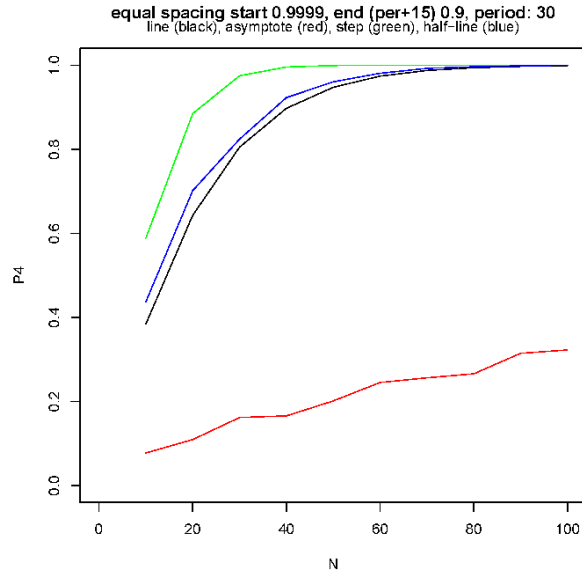
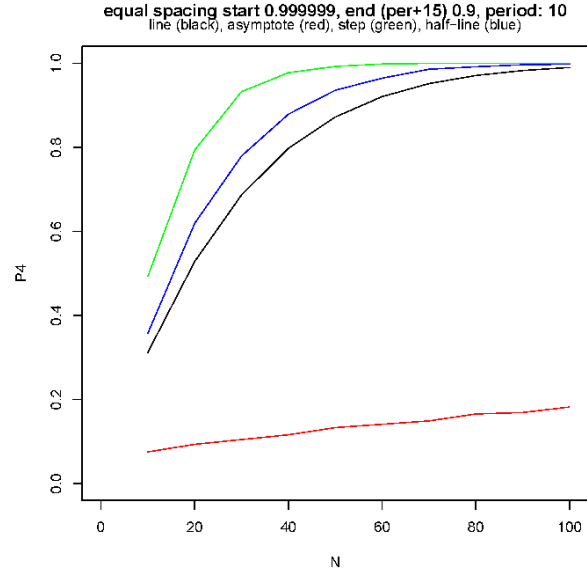
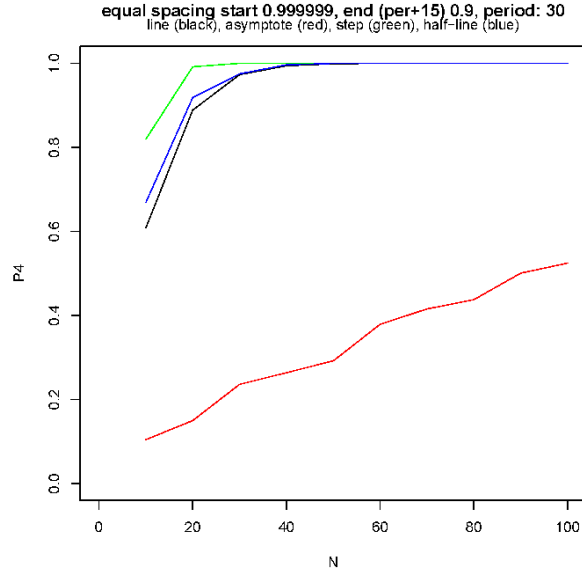


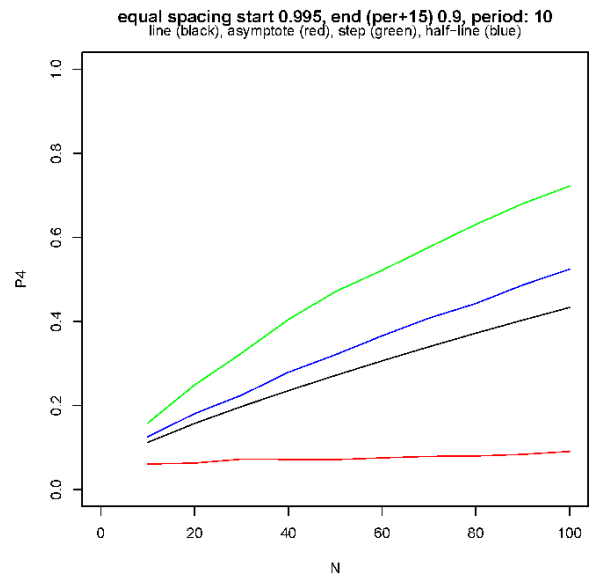
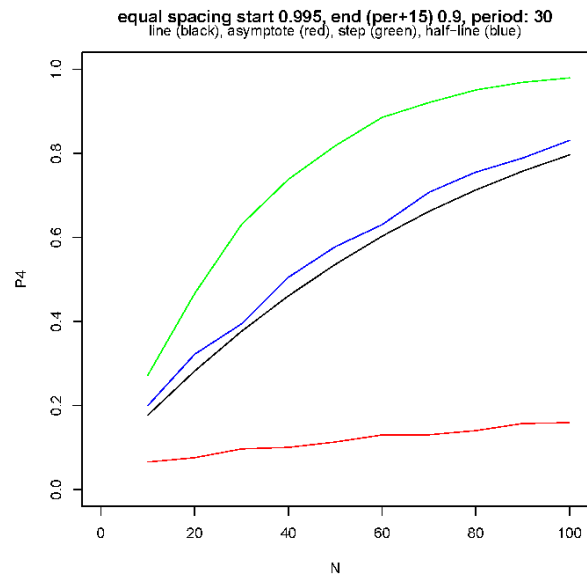
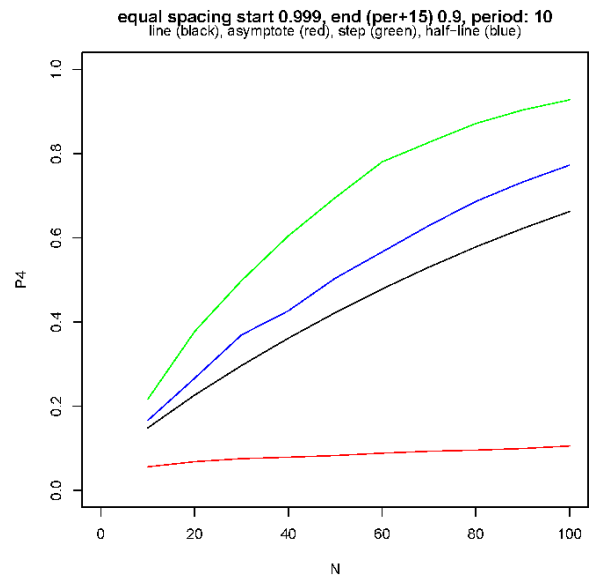
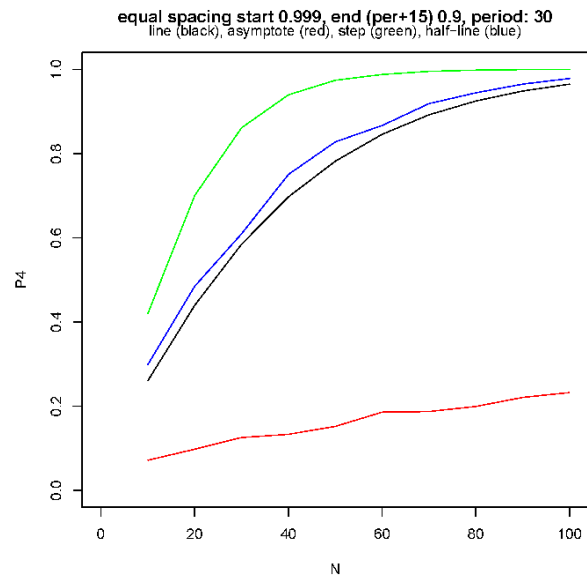
P4 Results, End=0.9999, Varying Start

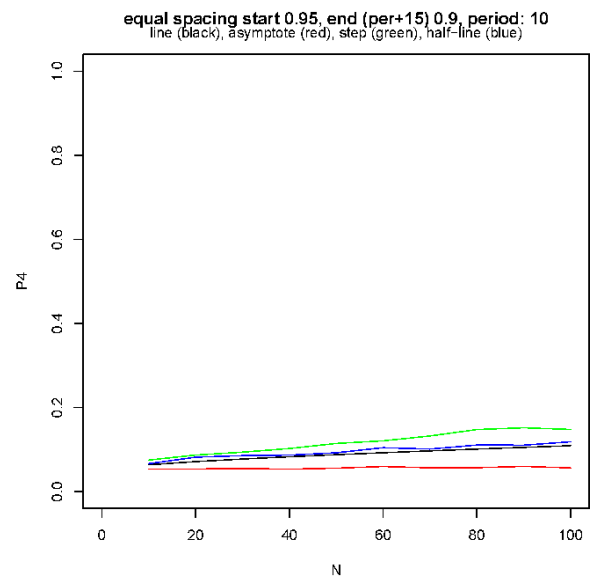
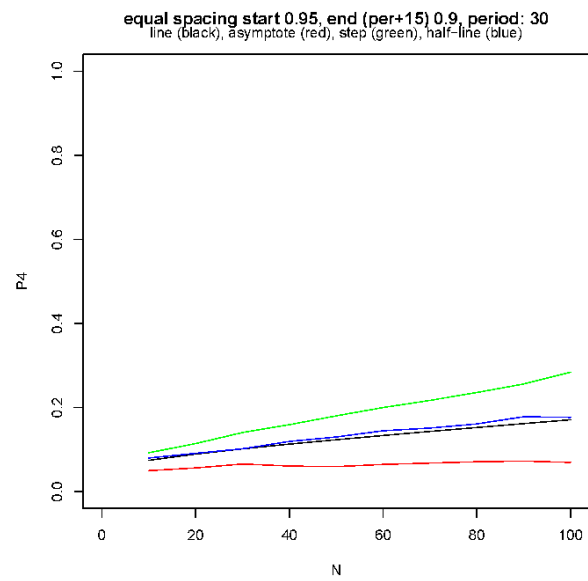
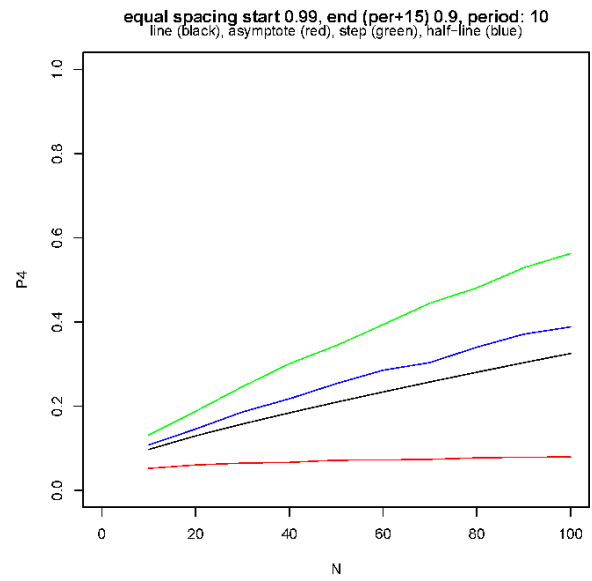
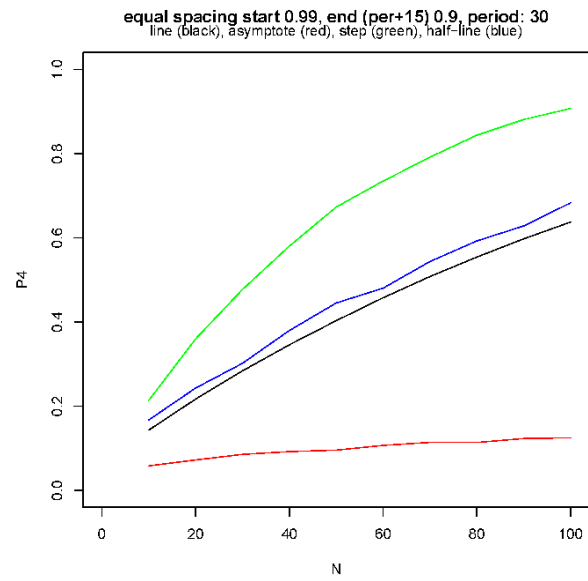


APPENDIX I: METRIC 4 SENSITIVITY STUDY GRAPHS, NON-LINEAR TREND

*P4 Results, Three non-linear trends (half-line, asymptote, step function)
End=0.90, Varying Start*







DISTRIBUTION

Distributed electronically

- 1 Michael Hamada, LANL
- 1 Aparna Huzurbazar, LANL
- 1 Charles Hills, LANL
- 1 Earl Whitney, LANL

- 1 Jeff Hagerty, LLNL
- 1 Bill McLean, LLNL
- 1 Alix Robertson, LLNL

- 1 K. V. Diegert, NNSA
- 1 R. Streit, NNSA

- 1 MS0629 C. A. Cruz, 2950
- 1 MS0634 B. J. Gomez, 2957
- 1 MS0634 D. R. Sherman, 2958
- 1 MS0635 L. M. Hickman, 2957
- 1 MS0829 J. M. Sjulín, 413
- 1 MS9007 R. L. Bierbaum, 8245
- 1 MS9007 R. M. Zurn, 8245

- 1 MS0899 Technical Library, 9536 (*electronic copy*)

